Acquisition automatique d'informations lexicales à partir de corpus : un bilan

Ronan Pichon, Pascale Sébillot

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Acquisition automatique d'informations lexicales à partir de corpus : un bilan

Ronan Pichon et Pascale Sébillot

N° 3321

Décembre 1997

THÈME 3

ISSN 0249-6399

Acquisition automatique d'informations lexicales à partir de corpus : un bilan

Ronan Pichon* et Pascale Sébillot*

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Repco

Rapport de recherche n 3321 — Décembre 1997 — 62 pages

Résumé : Dans ce rapport, nous faisons un bilan de multiples travaux qui voient actuellement le jour en acquisition automatique d'informations lexicales à partir de corpus. Nous les présentons en fonction de l'aspect du travail d'un lexicographe qu'ils peuvent assister, c'est-à-dire acquérir les termes d'un domaine afin de circonscrire le vocabulaire à intégrer à un lexique, structurer un lexique en classes sémantiques reliées entre elles, l'adapter à un langage, et déterminer les prédicats d'un domaine. Nous concluons en discutant ces diverses recherches dans la perspective d'une construction automatique de lexiques sémantiques, tout en dégageant certaines règles générales à suivre pour un tel travail.

Mots-clé : sémantique lexicale, acquisition d'informations lexicales, lexique sémantique, traitement de corpus.

(Abstract: pto)

* {rpichon}{sebillot}@irisa.fr

Unit´e de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
T´el´ephone : 02 99 84 71 00 - International : +33 2 99 84 71 00
T´el´ecopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Corpus-Based Approaches for Lexical Acquisition: an Overview

Abstract: This text is an overview of numerous works about corpus-based approaches for lexical acquisition. This presentation is directed by the lexicographer's tasks that these studies are supposed to facilitate, i.e. term acquisition in order to circumscribe the vocabulary of the lexicon of a given domain, structuring of a lexicon with linked semantic classes, adjustment to a language, and determination of the predicates of a domain. We discuss the integration of these researches into an automatic construction of semantic lexicons, and we point out some general rules for such a work.

Key-words: lexical semantics, lexical acquisition, semantic lexicon, corpus-based approaches.

Table des matières

1 Introduction

Traditionnellement, un lexique est défini comme l'ensemble du vocabulaire de la langue considérée et des propriétés linguistiques de ses éléments. L'élément de vocabulaire généralement considéré est le mot. Dans le cadre d'une application de traitement automatique de la langue naturelle[1], le lexique est restreint au vocabulaire du langage concerné[2] et il recense les informations sur les entités manipulées par l'application. Ainsi, par exemple, pour un correcteur orthographique sommaire, le lexique énumère l'ensemble des formes admises des mots du langage afin de pouvoir reconnaître si un mot donné est valide ou non.

Outre les formes admises des mots, un lexique se focalise historiquement sur leurs sens, qui peuvent être représentés de manières très différentes. En sémantique [Pot92, Ras96] et en lexicologie [Cru86], on fait la distinction entre le *signifiant* et le *signifié*. Le signifiant est le signe linguistique utilisé pour désigner le signifié. Dans le cadre d'une représentation lexicale, on considère que ce signifiant est le *lexème*[3]. Le signifié est ce qui est désigné ; en sémantique, on dit qu'il est le *concept* évoqué par une occurrence du signifiant ; en lexicologie, on dit qu'il est un des concepts[4] associés au lexème. Par exemple, dans la phrase « asseyez vous sur ce siège », le concept associé au mot siège est : OBJET FAIT POUR QUE L'ON S'ASSOIE DESSUS[5]. Par contre, au lexème **cellule**, on peut associer le concept ORGANISME VIVANT, mais également le concept CELLULE DE PRISON.

L'organisation la plus simple du lexique est une nomenclature[6] des lexèmes. Pendant longtemps, les lexiques utilisés dans les applications du TALN [Dow79,

1. Noté TALN.
2. Par la suite, on confond, sauf mention contraire, le lexique de la langue avec le lexique associé à une application du TALN.
3. Le lexème est communément défini comme étant l'entité morphologique référencée par le lexique. On choisit généralement de travailler au niveau du mot.
4. Dans le choix des concepts entre toujours une part d'arbitraire.
5. On notera que si on remplace *ce siège* par *cette chaise*, le concept évoqué sera le même.
6. La liste des mots du langage. C'est également la forme la plus ancienne. Les premiers lexiques connus, retrouvés sur des tablettes d'argile mésopotamiennes, se présentent sous cette forme.

Cru86, Rad88, Pus95, Weh97] ont été organisés sous cette forme. C'est aux éléments de cette nomenclature que sont alors attachées les informations dont l'application a besoin. Par exemple, dans une analyse syntaxique d'une phrase, le rôle de verbe doit être associé à un mot dont la catégorie grammaticale est VERBE. Si l'analyse désigne un complément d'objet direct à ce verbe, ce dernier doit en plus être reconnu comme transitif. Ces contraintes sont exprimées dans la nomenclature, où elles sont associées aux mots du vocabulaire. Le principal problème lié à cette approche est que chaque entrée lexicale est alors vue comme une entité autonome. Il est difficile de retrouver les relations, notamment sémantiques, entre entités lexicales, celles-ci étant représentées de façon implicite.

Or, une part importante du sens d'un mot se détermine par rapport à d'autres mots [RCA94, Pus95]. C'est pourquoi le lexique ne doit plus seulement être vu comme une suite de lexèmes et de définitions. Les données que référence le lexique sur les éléments du vocabulaire doivent en fait se situer à deux niveaux :

- *au niveau du lexème* : à ce niveau, on exprime ce que la linguistique nous apprend sur le mot. En sémantique lexicale, on s'attache à décrire quels concepts sont associés au lexème ;
- *au niveau des relations entre les lexèmes* : connaître les propriétés linguistiques d'un mot peut ne pas être suffisant ; il est généralement utile de pouvoir les relier avec celles d'autres mots. C'est pourquoi on s'attache à représenter les relations entre les entités du lexique. En sémantique, on cherche à mettre en évidence les mots dont les sens sont liés.

Un *lexique sémantique* est une structure qui met en évidence les relations sémantiques entre les lexèmes et les concepts, ainsi que celles qui lient les lexèmes entre eux et les concepts entre eux. Un couple composé d'un lexème et d'un concept qui lui est associé forme une *entrée lexicale*.

La première étape pour la construction d'un tel lexique[7] est de déterminer ces entrées lexicales, c'est-à-dire de trouver quels lexèmes sont associés à quels concepts. S'il est possible d'associer un même concept à plusieurs lexèmes, on

7. Si on suppose les lexèmes identifiés et les concepts déterminés.

définit une *classe sémantique* regroupant ces lexèmes. On peut ensuite définir des *traits sémantiques*[8] qui permettent de distinguer les lexèmes au sein d'une même classe. Par exemple, au concept OBJET FAIT POUR QUE L'ON S'ASSOIE DESSUS, on peut associer les lexèmes **chaise** et **fauteuil**, distinguables par le fait que **fauteuil** possède le trait sémantique /a des accoudoirs/ alors que **chaise** en est dépourvu. En général, on se sert de relations ontologiques[9] pour définir ces traits. Dans certains cas, il est possible d'utiliser ces traits sémantiques[10] pour organiser les éléments d'une classe sémantique donnée en une taxinomie, telle que celles utilisées en botanique ou en zoologie par exemple.

Une fois définies les relations au sein des entrées lexicales, c'est-à-dire entre concepts et lexèmes, on peut s'intéresser aux relations entre les entrées lexi-

8. Dans le choix des traits sémantiques, il y a, comme dans celui des concepts, une part d'arbitraire.

9. En intelligence artificielle en général et en représentation des connaissances en particulier, on définit une ontologie comme une représentation du monde ou d'un domaine de connaissances obtenue en structurant les éléments du monde ou les concepts du domaine. Or, les relations les plus fréquemment utilisées pour structurer de telles ontologies sont des relations classiques de sémantique lexicale. C'est pourquoi on confond souvent, et parfois abusivement, dans le domaine du traitement de la langue naturelle, lexique sémantique et ontologie. On qualifiera, dans cette étude, une relation d'ontologique si elle se définit par rapport au monde, ou plus formellement si elle fait appel aux caractéristiques des signifiés, par opposition aux propriétés des signifiants que nous qualifions ici de linguistiques. Par exemple, le fait qu'un fauteuil possède des accoudoirs alors qu'une chaise n'en possède pas est une propriété ontologique.

10. On note que ces traits sémantiques ont pour objet de différencier les éléments d'une classe sémantique et non de fournir une *définition* du concept associé à la classe.

cales. Ces relations sont dites *relations lexicales*[11]. Les relations lexicales sémantiques[12] les plus communément admises sont :

- *la polysémie*, qui désigne le fait qu'un même lexème est en relation avec plusieurs concepts au sein du lexique[13]. Ainsi **cellule** est un lexème polysémique ;
- *la synonymie*, qui existe quand deux lexèmes (ou plus) sont substituables dans le même contexte[14], c'est-à-dire que l'on peut remplacer en toute occasion un lexème par l'autre sans que le sens exprimé soit modifié. On parle également de synonymie partielle, si les deux lexèmes ne sont interchangeables que dans certains contextes[15], cas le plus fréquent ;
- *l'antonymie*, qui désigne une opposition entre les sens de deux entrées lexicales, par exemple **blanc/noir** ou **riche/pauvre** ;
- *l'hyperonymie*, qui désigne une relation entre entrées lexicales du plus général au plus spécifique. Ainsi **animal** est un hyperonyme de **chien**. On note également cette relation « sorte de »[16] ;
- *l'hyponymie*, qui est la relation réciproque de l'hyperonymie. Par exemple **chien** est un hyponyme de **animal**.

Il est donc possible, grâce à ces relations, de construire un *lexique sémantique*. Cependant, l'élaboration d'un tel lexique pour une application du TALN se heurte à un obstacle important : elle demande énormément de temps à un lexicographe [BP96]. De plus, une fois construit, il est souvent difficile d'utiliser

11. On considère donc que toute relation qui met en jeu une ou plusieurs entrées lexicales est une relation lexicale.

12. Est qualifiée de relation lexicale sémantique toute relation lexicale fondée sur le sens des entrées lexicales qu'elle lie.

13. En fait, on parle également d'*homonymie* si les concepts considérés n'ont rien en commun. Dans ce cas, on définit un nouveau lexème. Mais la frontière entre homonymie et polysémie est souvent floue. Par exemple, d'un dictionnaire à l'autre, une même relation sera qualifiée de polysémie dans l'un et d'homonymie dans l'autre.

14. On considère généralement les phrases ou les syntagmes comme contextes pour vérifier l'existence d'une relation de synonymie.

15. Par exemple : occuper/remplir une fonction et *occuper/remplir un formulaire.

16. IS-A en anglais.

le lexique pour une autre application. C'est pourquoi de nombreux travaux s'intéressent actuellement à l'*acquisition automatique d'informations lexicales*[17]. De telles informations peuvent être trouvées à l'aide de deux sources : les dictionnaires en ligne[18] et les corpus de textes[19]. Les dictionnaires sont intéressants car très riches. Leur utilisation dans le cadre d'une application d'acquisition lexicale est toutefois limitée par deux inconvénients : il est, d'une part, très difficile de retrouver les relations lexicales liant les entrées du dictionnaire et, d'autre part, les informations présentes ne sont pas toujours suffisantes pour obtenir le sens précis des lexèmes dans le cas d'une application spécifique à un domaine particulier. C'est pourquoi la majorité des travaux en acquisition lexicale travaille à partir de corpus. Ceux-ci présentent en effet l'avantage d'une plus grande souplesse : il est assez facile de réunir des textes traitant d'un domaine particulier, et on peut suivre l'évolution du langage utilisé dans le domaine étudié, par l'ajout de nouveaux textes à ces corpus.

Dans ce rapport, nous souhaitons faire un bilan des multiples travaux qui voient actuellement le jour dans le domaine de *l'acquisition lexicale*. De tels travaux permettent de mettre en évidence des informations susceptibles d'être représentées au sein d'un lexique. Nous sommes conscients de ne pouvoir présenter l'ensemble des travaux d'un domaine extrêmement productif. Nous essayons toutefois de dégager quelques grandes familles parmi cet ensemble afin d'avoir un aperçu relativement complet du domaine.

Il est possible de décomposer les travaux de ce domaine en fonction de l'aspect du travail du lexicographe qu'ils peuvent assister :

1. extraire les termes d'un domaine. L'extraction de terminologie se donne pour but de retrouver, dans un corpus, les termes du domaine. De tels travaux permettent de mettre en évidence les lexèmes que l'on recherche. Par exemple, si *complexité de l'algorithme* est reconnu comme un terme du domaine, alors *complexité* et *algorithme* seront deux lexèmes du lexique

17. Nous parlerons indifféremment d'acquisition d'informations lexicales ou d'acquisition lexicale.

18. Ces dictionnaires sont des versions numériques des dictionnaires classiques.

19. C'est-à-dire un ensemble de textes choisis pour représenter tout ou partie d'un langage. On notera que certains dictionnaires sont construits par une analyse manuelle d'un corpus de textes, comme le Trésor de la Langue Française.

du domaine. Cette famille ne relève pas vraiment de l'acquisition lexicale ; elle se situe avant tout travail sur le lexique proprement dit, mais permet de déterminer le vocabulaire du domaine ;

2. structurer le lexique. C'est le cœur du travail de construction du lexique. On peut distinguer plusieurs sous-familles parmi ces travaux :

   (a) mettre en évidence des classes sémantiques. Ces travaux se donnent pour but de regrouper les mots qui sont susceptibles d'exprimer des concepts proches au sein du corpus. Ces regroupements se font généralement en fonction du contexte syntaxique dans lequel ces mots apparaissent au sein du corpus étudié. Ceci permet de déterminer les concepts du domaine ;

   (b) mettre en évidence et typer les relations lexicales entre les concepts représentés par des classes sémantiques connues. Les relations étudiées peuvent être de type ontologique ou linguistique ;

3. mettre à jour les connaissances représentées au sein d'un lexique. Ici, on suppose l'existence d'un lexique que l'on doit adapter à un usage précis. De tels travaux ont deux objectifs :

   (a) enrichir le vocabulaire du lexique. Ceci se fait par l'ajout de nouveaux mots ou la découverte de classes sémantiques plus pertinentes que celles qui sont déjà utilisées ;

   (b) préciser les informations déjà présentes, par exemple en supprimant des interprétations peu crédibles de mots dans le domaine étudié. Par exemple, pour le lexème **souris**, on peut ne considérer que le concept OBJET MÉCANIQUE DE POINTAGE dans le domaine de l'informatique et le concept PETIT ANIMAL RONGEUR dans le cadre de la biologie ;

4. retrouver des informations sur les *prédicats*[20] du domaine, c'est-à-dire les mettre en évidence et déterminer leurs structures argumentales. Ces

20. Une fois déterminées les relations syntaxiques au sein d'une structure complexe, telle qu'une phrase ou un groupe nominal, on utilise fréquemment le formalisme de la logique du premier ordre pour en exprimer le sens. On parle alors de sémantique vériconditionnelle. L'un des éléments clés de ce formalisme est le prédicat. En sémantique, il est utilisé pour

prédicats, qui sont généralement associés aux verbes, sont fréquemment utilisés pour exprimer le sens d'une structure complexe (phrase, nom composé, etc.). Leur acquisition, même si ce n'est pas à proprement parler de l'acquisition lexicale, permet de bénéficier, outre la description des entrées lexicales des verbes auxquels ils sont liés, de données sémantiques très précises sur le domaine étudié.

Afin d'expliciter concrètement ce qu'est un lexique sémantique et comment il peut être utilisé, nous débutons ce rapport par une rapide présentation d'exemples de tels lexiques fréquemment utilisés en TALN. Nous exposons ensuite un ensemble de travaux d'acquisition lexicale qui représentent un aperçu des approches que nous venons d'énumérer. Lors de cette présentation, nous avons choisi de décrire plus en détail certaines des recherches que nous citons afin d'avoir une vue suffisamment précise des différentes méthodes qui sont utilisées dans ce domaine. L'exposé que nous faisons de tous ces travaux suit le plan de regroupement en familles donné ci-dessus.

2 Quelques exemples de lexiques

Dans cette partie, nous décrivons tout d'abord deux lexiques sémantiques d'usage courant dans des applications récentes du TALN. Le premier d'entre eux, WordNet [21] [MBF+90], est une base de données lexicales construite à l'aide

caractériser la relation qui lie les éléments de la structure étudiée ; il est associé à un élément lexical de la structure. Cet élément peut être identifié par une analyse syntaxique de la structure. On assimile la tête de la structure à un prédicat qui va chercher ses arguments parmi les autres éléments de la structure. Dans le cas d'une phrase, on associe généralement le verbe de la proposition principale et le prédicat de la phrase. Ainsi, la phrase : « (Les enfants) (aiment (tous (les jouets bruyants))).» peut s'écrire en logique du premier ordre sous la forme suivante : $\forall x, \lambda x.BRUYANT(\text{thème : x}) \Longrightarrow AIME(\text{agent : enfant, thème : x})jouet$. La représentation ainsi définie permet de montrer comment les éléments de la phrase se combinent pour former son sens. Les indicateurs thématiques, ici thème et agent, permettent d'expliciter le rôle des arguments des prédicats dans la construction de ce sens. De plus, les arguments des prédicats peuvent également être sémantiquement typés ; par exemple l'argument agentif du prédicat AIMER doit posséder l'attribut sémantique /humain/. Ceci permet de reconnaître certaines phrases telles que : *« les jouets aiment tous les enfants» comme n'ayant pas de sens.

21. WordNet est une marque déposée de l'université de Princeton.

des seules relations lexicales, et donc sans description formelle du contenu des entrées lexicales. Le second est le lexique génératif de J. Pustejovsky [Pus91, Pus95], qui définit un formalisme de représentation en sémantique lexicale. Nous nous limitons ici à ces deux exemples bien que d'autres lexiques tels que le *Longman Dictionary of Contemporary English*, COMLEX [GMM94], *ANLT* [BBC+87] et la classification établie par B. Levin [Lev93] pour les verbes anglais soient également d'un usage commun. Afin de comprendre comment un lexique sémantique peut être utilisé dans le cadre d'une application réelle, nous terminons en décrivant le lexique développé par C. Fabre et P. Sébillot [FS95, Fab96] pour un travail sur le calcul du sens de composés nominaux[22].

2.1 WordNet

WordNet est un lexique sémantique élaboré par des psycholinguistes à l'université de Princeton [MBF+90], qui peut être considéré comme un outil de référence. G. Miller et al. ont entrepris de bâtir un lexique autour des relations de sémantique lexicale que nous avons présentées précédemment, et WordNet a été construit manuellement en s'appuyant sur des travaux de psycholinguistique. Pour expliquer l'optique de ce lexique, ses auteurs le présentent en distinguant deux approches dans la représentation de la sémantique des entrées lexicales :

- *la représentation descriptive*, où l'on s'attache à donner une description complète du sens des concepts que l'on veut référencer. C'est l'approche la plus ambitieuse et la plus difficile à mettre en œuvre. Elle consiste à reconnaître les concepts exprimés par le langage étudié et à en donner une description formelle, par exemple définir ce que représente le concept OBJET FAIT POUR QU'ON S'ASSOIE DESSUS ;

- *la représentation différentielle*, où l'on cherche à mettre en évidence les concepts et leurs relations. C'est l'approche qui est choisie dans WordNet. Les concepts sont représentés par des *synsets*, ensembles de mots qui, dans un contexte donné, sont interchangeables. Un même mot peut appartenir à plusieurs synsets qui représentent ainsi la polysémie du mot

22. Plus précisément, des structures binominales complexes de la forme Nom Nom en anglais et Nom à/de (dét) Nom en français.

étudié. On s'attache donc plus à mettre en évidence les relations lexicales entre les concepts du lexique qu'à représenter ce qu'ils référencent, ce qui est brièvement fait en donnant un exemple de phrase où le concept est exprimé.

Dans WordNet, les entités lexicales[23] sont réparties en quatre catégories grammaticales : les noms, les verbes, les adjectifs et les adverbes ; elles sont organisées en quatre réseaux sémantiques, une par catégorie, dont les nœuds représentent les synsets et les arcs les relations lexicales entre ces synsets. Les relations considérées sont des relations classiques en sémantique lexicale. On peut ainsi déterminer quels sont les liens possibles entre mots ou entre synsets.

Voici, par exemple, quels sont les sens de **sand** dans la base WordNet :

```
Sense 1
sand
       => soil, dirt
--------------
Sense 2
backbone, grit, guts, sand
       => fortitude
--------------
```

Le premier sens associé à **sand** est celui de SABLE. C'est le sens le plus commun du mot, c'est pourquoi il apparaît en premier. Ce sens représente un concept hyponyme de celui qui est associé au synset **{soil, dirt}**. Il ne possède pas de synonyme. **Sand** est également associé à un deuxième concept, qui appartient à niveau de langage plus familier. Il signifie AVOIR DES TRIPES et possède plusieurs synonymes : **grit, guts** et **backbone**. Ce sens de **sand** est un hyponyme de celui associé à **fortitude**, à savoir FORCE DE L'ESPRIT.

La réserve principale qu'on peut émettre au sujet de WordNet[24] est que, dans sa volonté de couvrir toute la langue anglaise, il introduit des sens marginaux pour certains mots sans préciser la catégorie de langage qui utilise

23. C'est-à-dire les *synsets*.

24. Outre le fait d'être une base lexicale de l'anglais, défaut qui va perdre de son importance, puisque plusieurs projets dont le but est de construire un WordNet pour diverses langues européennes (le français, l'allemand, l'espagnol, l'italien, le néerlandais et l'estonien) ont vu le jour.

ces sens[25]. De plus, il existe au sein du lexique des ensembles de synsets qui sont structurés en ontologies locales[26] sans qu'il soit fait la distinction avec les relations lexicales classiques. Il en ressort que, pour utiliser les informations contenues dans la base lexicale dans une application, il faut souvent en choisir un sous-ensemble[27], ce qui pose le problème du choix du sous-ensemble approprié et donc d'une analyse des besoins de l'application.

2.2 Le lexique génératif

Le travail effectué par J. Pustejovsky autour de son lexique génératif [Pus95] procède d'une démarche différente. Ici, il n'est plus question d'isoler des ensembles de mots pour former des concepts, mais de décrire, pour chaque lexème, une entrée lexicale et la structure sémantique qui lui est associée à partir de relations lexicales et de propriétés sémantiques classiques. Cette approche, si on se réfère à la terminologie de G. Miller et al [MBF+90], est apparentée à une représentation descriptive. De plus, le formalisme utilisé permet de représenter, dans une seule entrée lexicale, l'ensemble des facettes polysémiques d'un lexème. Ainsi, chacun d'entre eux est défini par quatre attributs : une structure argumentale, une structure événementielle, une structure qualia, une structure décrivant les mécanismes d'héritage des attributs entre lexèmes. Le lexique, constitué par les lexèmes ainsi définis, est structuré en un treillis dont les arêtes sont déterminées par les relations lexicales entre les lexèmes. Précisons les attributs des lexèmes.

1. *La structure argumentale* énumère les différentes facettes sémiques du lexème, c'est-à-dire les différents concepts qu'on peut associer au lexème dans une interprétation, par exemple ANIMAL et VIANDE pour le lexème **agneau**. Cette structure recense également le type des entités qui sont susceptibles d'être associées au lexème considéré dans une interprétation, par exemple ANIMAL CARNIVORE pour x dans « x a mangé l'agneau ».

25. Comme cela se fait dans les dictionnaires par des mentions du type : familier, argot, etc.

26. Il s'agit, ici, de définir une structure locale pour des concepts qui sont significatifs au sein d'un domaine particulier tel que la zoologie.

27. C'est-à-dire ne conserver que certains synsets et certaines relations lexicales, ceux et celles qui sont utiles pour l'application considérée.

2. *La structure événementielle* exprime les types des événements associés au lexème, par exemple des actions qui peuvent changer l'état de l'objet référencé par le lexème (pour une maison, un de ces événements peut être sa destruction).

3. *La structure qualia* est elle-même composée de quatre champs. C'est dans cette structure qu'est représenté le sens du mot. On y exprime comment les arguments et les événements décrits précédemment se combinent dans les prédicats associés aux champs de la structure. À chacune de ces qualia correspond un type d'information sur le lexème :

   (a) *la qualia formelle* exprime ce qui distingue le lexème des autres. Elle joue un peu un rôle de typage ;

   (b) *la qualia constitutive* décrit les relations entre l'objet représenté par le lexème et ses constituants ;

   (c) *la qualia télique* définit ce à quoi sert typiquement l'entité représentée ;

   (d) *la qualia agentive* explicite comment l'entité est créée.

   Il n'est pas nécessaire que pour chaque lexème ces quatre champs existent. La structure qualia détermine comment interpréter le sens du lexème en contexte.

4. *La structure d'héritage* définit les mécanismes d'héritage entre les lexèmes.

Une entrée lexicale ainsi définie est un *paradigme lexical conceptuel* (en anglais, *lexical conceptual paradigm*, abrégé en *lcp*).

Les deux premières structures sont classiques en sémantique. Les deux dernières forment l'originalité du travail de Pustejovsky. La structure qualia permet d'exprimer les relations sémantiques potentielles entre plusieurs lexèmes, le typage des relations qui y sont référencées devant suffire à sélectionner celle ou celles effectivement possibles. Le mécanisme d'héritage définit les liens entre les lexèmes, c'est-à-dire qu'il permet à certains lexèmes d'hériter de tout ou partie de la structure qualia d'autres lexèmes.

Prenons l'exemple du lexème **newspaper** dont la représentation lexicale est détaillée ci-dessous :

$$\left[\begin{array}{lll} \textbf{newspaper} & & \\ \text{ARGSTR} & = & \left[\begin{array}{lll} \text{ARG1} & = & x : org \\ \text{ARG2} & = & y : info.physobj \end{array}\right] \\ \text{QUALIA} & = & \left[\begin{array}{lll} \textbf{org.info.physobj-lcp} & & \\ \text{FORMAL} & = & y \\ \text{TELIC} & = & read(e_2, w, y) \\ \text{AGENT} & = & publish(e_1, x, y) \end{array}\right] \end{array}\right]$$

Cette entrée lexicale présente trois aspects possibles du sens de **newspaper** : l'INSTITUTION (*org*), l'OBJET MATÉRIEL (*physobj*) et les INFORMATIONS CONTENUES DANS UN JOURNAL (*info*). Il est possible de décrire le sens représenté dans cette entrée lexicale de la façon suivante :

- la structure argumentale (ARGSTR) indique que, dans le processus d'interprétation, **newspaper** est vu soit comme une INSTITUTION, soit comme un OBJET **et** une INFORMATION ;
- la structure qualia indique :
  - qu'on utilise la structure qualia associée à l'archétype complexe **org.info.physobj-lcp**, les autres informations présentées dans la structure qualia étant spécifiques à **newspaper** ;
  - que **newspaper** est fait pour être lu. La lecture est un événement associé à l'aspect **info.physobj** du lexème. De plus, il faut un lecteur, ici w ;
  - que ce que représente le lexème est le résultat d'un événement, la publication, où **newspaper** est vu sous ses deux angles. En effet, le journal, vu comme un OBJET et une INFORMATION, est publié par l'INSTITUTION.

Les relations de sémantique lexicale entre les lexèmes permettent de définir un mécanisme d'héritage où chaque champ de la structure qualia peut hériter de différents *lcp*. Il est ainsi possible de bien rendre compte des relations

sémantiques entre les lexèmes. En fait, cela revient à définir quatre treillis, chacun étant consacré à un des champs de la qualia.

Les principales réserves qu'on peut exprimer sur ce formalisme[28] sont les suivantes :

1. la première, qui est liée au fait que Pustejovsky ait décrit un formalisme, porte sur la construction d'un lexique de grande taille utilisant cette représentation. Dans [PAB93], l'auteur s'est livré à une tentative d'acquisition automatique d'informations sur corpus, mais le bilan n'est pas satisfaisant. Seules quelques pistes émergent de ces travaux ;

2. la seconde concerne les qualia. En effet, si Pustejovsky les dit suffisantes, on ne peut en être sûr, sauf à construire un lexique d'ampleur suffisante pour le prouver. De plus, la limite entre ce qui relève d'un aspect de la qualia et non d'un autre n'est pas toujours très claire.

Par contre, le principe d'une structure à plusieurs champs, conjugué avec les mécanismes d'héritages multiples et de typage complexe des lexèmes paraît très prometteur, le tout étant de définir une structure de type qualia adaptée à l'usage pour lequel on construit le lexique.

2.3 Une représentation lexicale dédiée

Pour illustrer comment un lexique sémantique peut répondre aux besoins d'une application réelle, nous terminons cette section en expliquant la représentation lexicale utilisée par P. Sébillot et C. Fabre pour leur modèle d'interprétation automatique de la sémantique des noms composés[29] anglais et français [FS95, Fab96]. L'objectif de ce travail est de déterminer l'ensemble des interprétations possibles de composés nominaux à partir d'une analyse de la sémantique de leurs constituants. Nous ne présentons ici que l'aspect de leur travail qui concerne la représentation lexicale des noms et l'utilisation qui en est faite dans le calcul du sens des composés étudiés.

28. Le lexique génératif, au contraire de WordNet, n'existe pas en tant que base de données lexicales.

29. Plus exactement, il s'agit de structures binominales complexes de type N N en anglais et N à/de (dét) N en français.

Dans leur modèle de calcul d'interprétations, les auteurs utilisent un prédicat pour exprimer un sens possible d'un nom composé. Ainsi, la sémantique de *seasickness pill* est représentée sous la forme : seasickness pill = prédicat : *heal*(instrument : *pill*, thème : *seasickness*). Le prédicat utilisé est déterminé par une analyse du sens des éléments du composé. Le fait qu'une pillule serve à guérir et que le mal de mer soit une maladie suffit à prédire que le composé peut signifier qu'il s'agit d'une pillule qui guérit le mal de mer. Pour associer un prédicat à une entrée lexicale ou à une classe sémantique, les auteurs utilisent les travaux de Pustejovsky [Pus95]. Pour simplifier, on peut considérer que les prédicats qui apparaissent dans la structure qualia d'une *lcp* sont ceux qui sont associés aux noms et classes sémantiques du lexique de l'application.

L'un des objectifs du modèle est de n'être pas lié à un domaine. C'est une des raisons pour lesquelles WordNet a été choisi pour servir de base à la représentation lexicale utilisée. Le lexique utilisé dans l'implémentation du modèle général ne contient que les synsets de WordNet qui représentent les concepts les plus généraux. Les classes sémantiques ainsi disponibles permettent de formuler des restrictions sur les arguments des prédicats. Par exemple, guérir nécessite un agent du type MÉDICAMENT et un thème du type MALADIE. Cette arborescence permet également d'associer aux noms les prédicats qui leur sont liés. Pour cela, ces prédicats sont intégrés dans le lexique, soit au niveau des entrées lexicales des noms, soit à celui des classes sémantiques de la hiérarchie WordNet.

Nous avons présenté quelques exemples de représentations lexicales existant dans le domaine de la sémantique lexicale. Nous avons pu voir qu'il est possible de mettre en évidence de nombreuses propriétés sémantiques au niveau du lexique. Comme dans le cas décrit ci-dessus, la plupart des applications du TALN n'utilisent en fait qu'une partie de ces informations. Une des raisons principales de ce choix réside dans la difficulté de construire ou d'adapter manuellement un lexique pour les besoins d'une application réelle. Ceci plaide fortement pour l'acquisition automatique de ces informations lexicales.

3 Acquisition lexicale

Dans cette section, nous nous intéressons à un ensemble de travaux d'acquisition de connaissances en corpus qui mettent en évidence des informations susceptibles d'être représentées au sein d'un lexique. Nous verrons d'ailleurs que quelques-uns d'entre eux sont parfois à la limite de l'acquisition lexicale proprement dite, mais en sont suffisamment proches pour illustrer notre propos sur ce domaine.

La richesse des informations sur les mots extractibles de corpus peut être illustrée par un article de G. Grefenstette [Gre94] qui définit trois niveaux (croissants) de relations sémantiques pouvant être mises à jour sur des textes :

1. les relations du *premier ordre* désignent les mots typiquement associés dans les textes étudiés. L'un des critères les plus utilisés pour détecter de telles relations est l'information mutuelle entre deux mots. Elle permet de mettre en évidence les couples de mots dont l'association dans le corpus diffère de celle attendue par une distribution aléatoire. On fait également usage de la fréquence de cooccurrences des mots dans le corpus pour détecter une relation du premier ordre. On peut dire que deux termes associés par une telle relation participent chacun au contexte de l'autre dans le corpus. De telles relations peuvent notamment servir à mettre en évidence la structure argumentale de prédicats ;

2. les relations du *deuxième ordre* sont celles qui unissent les termes apparaissant dans un contexte similaire, par exemple les noms qui apparaissent en argument des mêmes verbes. Ce sont des éléments partageant les mêmes relations du premier ordre. On peut s'en servir pour établir une classification des termes étudiés. Par contre, il est impossible de rendre compte de la polysémie des mots avec de telles relations. En effet, les termes ne peuvent appartenir qu'à une seule des classes ainsi calculées. En général, les classes formées sont des ensembles de mots candidats à être reliés entre eux par des relations lexicales, que celles-ci aient un caractère linguistique ou ontologique ;

3. les relations du *troisième ordre* sont les relations définissant, pour chaque terme, plusieurs associations, afin de représenter ses diverses facettes. De

cette façon, il est possible d'avoir, par exemple, une classe de noms contenant les objets manufacturés, dont le couteau, et une classe contenant les objets qui coupent, dont le même couteau. Outre la polysémie des termes, ceci permet de déterminer des relations lexicales entre ensembles de lexèmes.

Ces trois niveaux de représentation offrent une référence si on souhaite catégoriser les travaux d'acquisition lexicale en fonction du contenu des relations qu'ils mettent en évidence. Pour les relations du premier ordre, il s'agit d'associations au sein de structures syntaxiques, telles que la structure argumentale d'un prédicat, les groupes nominaux ou les noms composés[30]. Elles permettent de mettre en évidence des relations syntaxiques et sémantiques entre des éléments de catégories grammaticales diverses. Les relations du second et du troisième ordres servent quant à elles à mettre à jour des liens qui se retrouvent au sein du lexique dans une même catégorie.

Les travaux actuels en acquisition lexicale ne prétendent pas suivre entièrement la démarche d'un lexicographe dans son travail ; ils peuvent cependant tous l'accompagner et l'aider à certaines étapes de la constitution d'un lexique. Nous avons donc choisi de présenter les recherches menées dans ce domaine selon la tâche du lexicographe qu'elles peuvent assister.

Nous débutons cet exposé par des travaux d'extraction de terminologie, qui permettent de dresser une nomenclature des termes (généralement complexes) d'un domaine. Dans une optique de constitution automatique de lexiques, ces études peuvent servir à déterminer les lexèmes pertinents.

Nous nous intéressons ensuite à des travaux dont le but est de regrouper des mots en ensembles, selon les "utilisations" qui en sont faites dans des corpus. Après une validation experte, ces ensembles peuvent, par exemple, servir à déterminer les concepts exprimés par les mots du vocabulaire considéré dans le domaine étudié. Certaines recherches exposées montrent également comment retrouver des relations lexicales entre les ensembles obtenus. Nous verrons que ces relations sont caractérisées par deux ensembles de propriétés, syntaxiques

30. Ces deux dernières structures sont intéressantes car c'est généralement sous cette forme que sont exprimés les concepts les plus spécifiques d'un domaine.

et sémantiques d'un côté, ontologiques de l'autre, suivant qu'elles concernent respectivement les signifiants ou les signifiés.

Nous poursuivons en montrant comment il est possible de mettre à jour les informations contenues dans un lexique pour un usage particulier, en ajoutant, en éliminant ou en précisant des entrées lexicales parmi celles qui existent déjà.

Nous concluons par des travaux dont l'objectif est de retrouver les prédicats d'un domaine et leurs structures argumentales.

La plupart des travaux présentés dans cette section ne se limitent en fait pas à un seul de ces aspects. Nous avons toutefois choisi de les expliciter suivant celui qui semble le plus intéressant pour notre propos.

3.1 Extraction de terminologie

Les travaux qui suivent portent sur l'extraction de terminologie. Nous avons déjà dit que ce n'est pas vraiment de l'acquisition lexicale, le but commun de tous ces travaux étant de dresser une nomenclature des termes d'un domaine à partir d'un corpus de textes[31] portant sur le domaine en question. Ce seul objectif permet de construire une liste des lexèmes utiles si on souhaite établir un lexique du vocabulaire du domaine. Nous verrons cependant que ces travaux ne se limitent pas à ce rôle et permettent de mettre à jour des informations plus riches. En effet, une terminologie n'est pas seulement une nomenclature de termes, c'est également un ensemble de relations entre les éléments de cette nomenclature, même si celles-ci ne sont pas nécessairement des relations lexicales classiques.

3.1.1 Travaux de Béatrice Daille

Dans ses travaux, B. Daille [Dai94] s'intéresse au problème de l'acquisition de termes représentés sous la forme de noms composés[32]. Pour ce faire, elle définit un ensemble de patrons syntaxiques dont les noms composés recherchés respectent la forme, par exemple : N1 (prep (det)) N2 (*travaux de génie, point relais*). Ces patrons sont basés sur des résultats de linguistique sur la

31. Voire éventuellement à partir d'une nomenclature terminologique initiale.

32. Le terme nom composé est vu ici dans un sens très large, celui de structure complexe de type nominal.

composition nominale. Les prérequis des travaux menés sont donc de disposer d'un analyseur syntaxique susceptible de reconnaître ces patrons. Ceci se fait à l'aide d'automates qui recherchent les segments de phrases adéquats.

L'auteur détermine ainsi une liste de noms composés candidats à être des termes du domaine. Pour effectuer un tri dans cet ensemble et ne conserver que les candidats les plus vraisemblables, elle étudie un vaste ensemble de critères statistiques. Son but est de déterminer si les mots associés dans un composé sont liés dans une relation du premier ordre. De ces critères, ceux qui sont basés sur la fréquence semblent obtenir les résultats les plus pertinents.

Le principal intérêt de ce travail, pour notre étude, est cette analyse des critères d'association. Elle montre bien que le choix du critère est déterminant pour la qualité des résultats obtenus.

3.1.2 Travaux de Didier Bourigault et al.

Lexter[33] [Bou94] est un logiciel dont le but est tant d'acquérir les termes d'un domaine, que d'aider le cogniticien chargé de travailler sur cette terminologie dans l'organisation des relations entre les termes. Pour ce faire, Lexter extrait d'un corpus[34] de textes un ensemble de groupes nominaux candidats à être des termes du domaine que représente le corpus. Ces candidats termes sont repérés dans les textes par des marqueurs *frontières* tels que les verbes ou les conjonctions. Les groupes nominaux ainsi reconnus forment des entités complexes, appelées groupes nominaux maximaux, par exemple *décision de renforcement des réseaux régionaux*. Lexter les décompose ensuite en couples (tête, expansion) par une analyse syntaxique des dépendances au sein du groupe nominal. Quand les têtes et expansions ainsi déterminées forment des groupes nominaux, elles sont décomposées de la même manière. On forme ainsi un réseau, dit terminologique, où chaque terme est relié à sa tête, son expansion, ainsi qu'aux termes dont il est la tête ou l'expansion.

Le résultat obtenu demande à être affiné par un terminologue pour être utilisé. Les termes extraits ne sont que des candidats et doivent être validés. De plus, les relations exprimées dans le réseau terminologique sont de type

33. Pour Logiciel d'EXtraction de TERminologie.

34. Le corpus doit au préalable être l'objet d'un étiquetage qui affecte à chaque lexème sa catégorie grammaticale.

syntaxique et ne suffisent pas à définir des associations conceptuelles. Dans ce but, Bourigault et al. proposent dans [AB96] une méthode de typage des termes permettant de déterminer ceux qui désignent les objets les plus génériques du domaine[35] et ceux qui désignent les attributs ou les actions les plus génériques. Le type est déterminé par une analyse statistique fondée sur la fréquence du terme parmi les têtes ou parmi les expansions dans le réseau terminologique[36]. Les attributs et actions sont ensuite discriminés par des critères linguistiques élaborés par D. Garcia et A. Jackiewicz [GJ95]. Ceci permet de typer une partie des termes.

Bourigault et al. utilisent ensuite les relations du réseau terminologique pour étendre ce typage à l'ensemble des termes par quelques règles très simples, par exemple : *si t désigne un objet, alors tous les termes dont il est la tête désignent des objets ; ceux-ci sont considérés comme les fils de t dans une relation « sorte-de »*. Ainsi, *ligne souterraine* est un fils de *ligne*. On définit donc, pour chaque objet du domaine, ses attributs, les actions qui peuvent s'appliquer à lui, ainsi que les termes qui désignent des objets plus spécifiques.

On peut enfin effectuer une classification des adjectifs et des noms du domaine, en regroupant ceux qui apparaissent dans des contextes similaires dans le réseau terminologique. Cette étape est détaillée dans [AB95, Ass96, Ass97] et est présentée plus loin dans ce rapport (cf. 3.2.3, page 36.).

3.1.3 Travaux de Christian Jacquemin

Les travaux de C. Jacquemin [Jac97] sur FASTR[37], logiciel d'acquisition de terminologie, ont pour objectif de retrouver, sur un corpus, des termes spécifiques d'un domaine et d'enrichir une terminologie préexistante. Pour ce faire, il recherche dans le corpus des variations morpho-syntaxiques de termes déjà connus.

35. C'est-à-dire, ici, des classes d'objets.

36. L'hypothèse est que plus la fréquence d'apparition d'un terme parmi les têtes est importante relativement à celle qu'il a dans les expansions, plus il est probable qu'il désigne un objet du domaine. Pour les actions et les attributs, on inverse le rôle des têtes et des expansions.

37. Pour Filtrage et Acquisition Syntaxique de TeRmes.

L'auteur définit un ensemble de règles de méta-grammaire pour reconnaître les variations de termes, par exemple reconnaître *vertèbres cervicales et dorsales* à partir du terme connu *vertèbre cervicale*. Ceci permet également de dire que *vertèbre dorsale* est un nouveau terme du domaine. De plus, il est possible, dans certains cas, d'utiliser ces variations pour retrouver des relations entre les mots, ici *dorsales* et *cervicales* qui désignent des concepts apparentés dans le domaine. Il est également possible avec son travail de mettre en correspondance deux formes différentes d'un terme, par exemple *frequency shift* et *frequency and it's shift*.

Ces travaux, dont la liste est évidemment loin d'être exhaustive, présentent un ensemble d'outils qui peuvent aider un terminologue dans son travail. Bien qu'ils ne traitent pas à proprement parler d'acquisition lexicale, ils peuvent tout de même participer à la démarche d'obtention d'informations sémantiques sur les termes d'un domaine. On peut envisager d'appliquer ces techniques pour aider à la construction d'un lexique spécialisé, pour déterminer le vocabulaire couvert par ce lexique, voire mettre en évidence l'existence de certaines relations entre les lexèmes ainsi identifiés.

3.2 Structuration du lexique

L'ensemble des travaux que nous présentons ici a pour objet de déterminer des informations susceptibles d'aider à la structuration d'un lexique sémantique. Ces études forment le cœur d'une démarche de construction d'un lexique à partir d'un corpus. Nous avons choisi de les regrouper en fonction de l'objectif que chacune d'entre elles cherche à atteindre :

1. regrouper des mots. Ces regroupements se font sur des critères différents suivant que l'ensemble lie des mots associés à un même autre mot [Gre93a], à un même concept [RS97] ou dans une classe sémantique [Aga95] ;

2. préciser une classification. Il s'agit de déterminer une classification sémantique sur un ensemble de mots en s'appuyant sur la connaissance d'une classification préexistante mais insuffisante. On distingue des tra-

vaux se situant dans le cadre du lexique d'un langage spécialisé [VFP91] ou dans celui de la langue [Bui97] ;

3. construire une ontologie. Ces travaux se proposent de structurer les mots d'un langage spécialisé pour représenter les relations qui regroupent les signifiés qui leur sont associés dans la terminologie du domaine étudié. Il est possible de mettre en évidence des ensembles de mots qui représentent des concepts similaires dans le domaine [Ass97, ZBHN97] ou de retrouver des relations sémantiques entre les mots qui ne soient pas seulement des relations lexicales [MPRT97] ;

4. mettre en évidence une relation linguistique. On peut s'intéresser aux relations entre un prédicat et ses arguments [Res93], ou bien chercher à mettre en évidence les classes sémantiques associées par une relation syntaxique donnée [BPV92] ou les indicateurs syntaxiques de relations lexicales [Mor97].

3.2.1 Regroupement de mots

Nous présentons ici des travaux dont l'objectif est de déterminer, à partir d'une analyse sur corpus, des ensembles de mots. Cette dénomination est volontairement vague, car elle recouvre des travaux très divers. On peut y distinguer deux grandes familles, selon ce que les ensembles ainsi trouvés prétendent représenter :

– des ensembles de mots candidats à être reliés entre eux par des relations lexicales. Ces regroupements sont les plus simples à obtenir et les classes ainsi formées sont plus à leur place dans un thesaurus que dans un lexique sémantique. En effet, on cherche à rassembler des mots qui sont associés à un même mot [Gre93b, Gre93a] ou à un même « concept[38] » [RS97] au sein du corpus étudié. Les critères qui déterminent si une association ainsi obtenue est valide peuvent être très souples. Ces travaux sont à la limite de l'acquisition de terminologie vue précédemment car ils ne

38. Le mot *concept* est à prendre, ici, dans un sens relativement général. Par exemple, au concept VEHICULE, on peut associer les mots qui désignent des véhicules, mais aussi ceux qui expriment un déplacement, une propulsion, etc.

fournissent, en fait, que très peu d'informations explicites sur les liens existant entre les éléments d'un groupe donné ;

– des ensembles de mots formant des classes sémantiques [Aga95]. Les prérequis sont ici plus importants et il est notamment nécessaire de valider les classes obtenues de façon plus sévère que celles obtenues par l'approche précédente.

Gregory Grefenstette [Gre93b, Gre93a]

Il est possible, pour reprendre l'analyse de G. Grefenstette [Gre93b], de caractériser les paramètres utilisés pour juger de la parenté des mots à classifier par la méthode selon laquelle ils ont été obtenus. Ces méthodes sont de deux types, selon que ces paramètres ont été déterminés par une analyse syntaxique ou par une analyse basée sur les fenêtres[39]. Elles peuvent être définies de la façon suivante :

1. dans une analyse basée sur la syntaxe, le contexte utilisé pour caractériser un mot est défini par la relation syntaxique existant entre le mot et ce contexte. Cela signifie que pour pouvoir typer, par exemple, les arguments d'un prédicat verbal, on extrait les mots désignés comme tels par un analyseur syntaxique. Le fait que les relations recherchées soient généralement simples permet d'utiliser des analyseurs syntaxiques dits de surface[40], puisqu'il n'est pas nécessaire de connaître toutes les dépendances syntaxiques entre les éléments des phrases étudiées ;

2. dans une analyse basée sur les fenêtres, le contexte est défini non plus par des relations syntaxiques entre les mots, mais par les mots les plus proches dans le corpus étudié. C'est ainsi que pour rechercher si deux verbes ont une structure argumentale similaire, on compare, par exemple, les n mots qui les suivent dans les phrases où on les rencontre dans le corpus. Cette méthode a pour principal avantage de pouvoir se passer de toute phase d'analyse syntaxique et, de ce fait, est plus facile à mettre en œuvre.

39. Libre traduction du terme anglais : « window-based ».

40. C'est-à-dire, des analyseurs qui ne mettent en évidence que certaines relations syntaxiques [Bri94].

La distinction se fait donc selon la façon dont sont déterminés les paramètres associés à un mot, mais ceci ne présage en rien de la méthode utilisée pour regrouper les mots à l'aide de ces paramètres. D'ailleurs dans [Gre93b], Grefenstette détermine les associations de termes à l'aide des mêmes critères statistiques sur ces paramètres. Pour ce faire, il regarde si les associations obtenues à partir des paramètres déterminés par l'une et l'autre méthode sont susceptibles d'être validées par des lexiques de référence. Les catégories ainsi obtenues ne forment pas des classes sémantiques, mais plutôt des ensembles de mots entre lesquels il est possible de retrouver des relations de sémantique lexicale[41]. Le bilan de cette comparaison est nuancé suivant que les noms sur lesquels il travaille se rencontrent fréquemment ou pas dans le corpus utilisé. Pour les mots les plus fréquents[42], qu'il est possible d'assimiler aux termes les plus importants du langage utilisé, les paramètres définis syntaxiquement permettent d'obtenir les meilleurs résultats. Pour les éléments les plus rares, le verdict s'inverse. Il semble donc, d'après l'auteur, que dans le cas où on s'intéresse à la terminologie d'un domaine, les paramètres utilisés pour analyser les termes gagneraient à être définis par une analyse syntaxique du corpus, de préférence à une analyse basée sur des fenêtres de mots.

Dans [Gre93a], l'auteur présente une méthode pour générer automatiquement un thesaurus spécialisé à partir d'une analyse sur corpus. Ce thesaurus recense les termes ayant un rapport avec les noms rencontrés dans le corpus. Ces travaux ne nécessitent pas de grandes connaissances lexicales a priori, seule la catégorie grammaticale des mots devant être connue. Les relations mises en évidence au sein du thesaurus peuvent notamment servir comme source d'informations pour contruire un lexique sur les termes du domaine. On verra toutefois qu'elles ne sont pas suffisantes.

Pour chacun des noms les plus fréquents du corpus, on obtient les informations (relations) suivantes :

- les noms rencontrés dans des contextes similaires au sein du corpus,
- les verbes dont le nom étudié est typiquement le sujet ou l'objet,

41. On pourra ainsi trouver au sein d'une même catégorie des synonymes, des antonymes, etc.

42. Dans son expérience, les 600 noms les plus fréquents du corpus.

- les noms composés où le nom étudié apparaît, ainsi que les noms composés apparaissant dans des contextes similaires,
- les mots apparaissant typiquement dans les mêmes textes que le nom étudié. Dans cette catégorie, on retrouve fréquemment des variantes morphologiques du nom considéré.

La méthode suivie pour déterminer les mots associés à chacun des noms étudiés est décrite ci-dessous.

1. Après une analyse lexicale [43] du corpus, les phrases subissent une analyse syntaxique afin de déterminer, pour chaque nom, les attributs [44] suivants :
   (a) les adjectifs auxquels il est associé,
   (b) les verbes dont il est le sujet et l'objet,
   (c) les noms qui le modifient, par apposition ou au sein d'un groupe prépositionnel.

   L'ensemble de ces attributs permet de déterminer, pour chaque nom, un vecteur d'attributs.
2. On calcule ensuite les similarités entre ces vecteurs en utilisant un coefficient de Jacquard.
3. On détermine ensuite que deux noms sont associés s'ils apparaissent chacun parmi les dix noms les plus fortement associés à l'autre.
4. Les verbes associés à un nom sont simplement ceux qui apparaissent le plus souvent parmi les attributs déterminés précédemment.
5. Déterminer les noms composés où le terme étudié apparaît est très simple. Par contre, pour découvrir les noms composés similaires à ceux qu'il vient de trouver, l'auteur se livre à une mesure de similarité analogue à

43. On considère ici l'analyse lexicale comme le processus qui associe à chaque mot du corpus sa catégorie grammaticale, verbe, nom, etc.

44. Au sens de paramètres utilisés pour retrouver les associations sémantiques entre les noms du corpus. Il ne s'agit pas ici d'attributs sémantiques.

la précédente, à la différence toutefois que les attributs utilisés ne sont plus déterminés par des dépendances syntaxiques mais par l'ensemble des mots composant les phrases où ils apparaissent. Ce dernier choix est lié au fait que ces termes sont trop peu nombreux au sein du corpus et que des travaux du même auteur, que nous venons de mentionner [Gre93b], ont montré qu'une telle approche est plus pertinente pour les termes peu fréquents dans un corpus.

Le thesaurus ainsi construit est d'autant plus pertinent que le terme considéré est fréquent au sein du corpus étudié. Cependant, le résultat présenté est encore insuffisant pour pouvoir être utilisé directement. Notamment certains noms importants n'apparaissent pas car leur fréquence est trop faible dans le corpus. De plus, les relations sémantiques qui lient un nom et les termes qui lui sont associés ne sont pas spécifiées ; c'est un travail à effectuer à la main. Ce thesaurus, s'il n'est pas directement exploitable, semble fournir une base solide pour constituer une représentation lexicale différentielle (cf. 2.1, page 11) des noms les plus fréquents du corpus.

Ellen Riloff et Jessica Shepherd [Ril93, Ril96b, Ril96a, RS97]

L'ensemble de ces travaux se situe dans le cadre de l'extraction d'informations telle qu'elle est pratiquée dans les études présentées lors des conférences MUC, rencontres annuelles où les chercheurs ont l'occasion de confronter leurs systèmes d'extraction d'informations à partir de corpus. Les travaux sont évalués suivant une procédure de test identique pour tous. Les chercheurs doivent développer un système qui extrait le maximum d'informations sur quelques types d'entités à partir de textes relativement courts. Par exemple, dans les conférences MUC-3 et MUC-4, le thème général était le terrorisme, et l'objectif que devaient atteindre les systèmes en compétition était d'extraire de dépêches d'agences de presse le plus d'informations possibles sur les actes de terrorisme qui y étaient décrits[45]. On fournit aux compétiteurs des squelettes de descriptions des entités à étudier, et leurs systèmes doivent repérer dans les textes les informations qui les concernent. Ils disposent également d'un corpus à l'aide duquel ils peuvent construire un outil adapté à l'extraction des informations

45. Par exemple identifier les victimes, savoir de qui elles étaient victimes, comment elles avaient été agressées (bombe, couteau), etc.

requises.

Les travaux de E. Riloff et al. montrent comment déterminer quelles parties des phrases (syntagmes ou mots) de textes sont liées à un concept donné, en utilisant des connaissances préalables de plus en plus restreintes.

Dans [Ril93, Ril96b], E. Riloff présente le logiciel AutoSlog, dont le but est d'épauler CIRCUS, l'outil d'extraction d'informations qu'elle utilise au sein d'un projet concourant aux conférences MUC. CIRCUS reconnaît les événements du domaine et les entités auxquelles ils se rapportent en utilisant un dictionnaire qui recense des patrons syntaxico-sémantiques d'événements, tels que *sujet verbe-au-passif* auquel est associée l'interprétation $[x]_{victim}$ *was murdered*. Or, la construction d'un tel dictionnaire d'événements porteurs d'informations prend beaucoup de temps et est une tâche répétitive, donc potentiellement automatisable. AutoSlog a pour but de partir de patrons uniquement syntaxiques et d'un étiquetage "conceptuel" des mots importants des textes, et de générer automatiquement les patrons syntaxico-sémantiques dont a besoin CIRCUS. Il fonctionne de la manière suivante :

1. on définit un ensemble d'heuristiques[46] qui décrivent des structures syntaxiques porteuses d'information, par exemple *sujet verbe-au-passif*, où l'on considère que le verbe apporte de l'information sur le sujet ;

2. on construit un corpus d'apprentissage où les groupes nominaux qui représentent les éléments importants à étudier sont identifiés comme tels et typés suivant les catégories "conceptuelles" (victime, etc.) sur lesquelles se fait l'extraction[47]. Ces catégories sont déterminées à l'avance par un expert ;

3. pour chaque groupe ainsi repéré, on vérifie s'il répond à une des heuristiques définies précédemment. Si oui, AutoSlog ajoute l'instance de la structure syntaxique à la liste des candidats à être des concepts du domaine. Par exemple, de $[The\ mayor]_{victim}$ *was killed by ...* et de l'heu-

46. Entre 13 et 15 heuristiques sont utilisées au cours de l'avancement des travaux.
47. Dans le cas du terrorisme, on identifie la ou les victimes de l'acte, les exécutants, etc.

ristique basée sur la règle *sujet verbe-au-passif*, on obtiendra le candidat concept : $[x]_{victim}$ *was killed* ;

4. l'ensemble des candidats proposés par le système doit ensuite être inspecté par un humain[48] pour rejeter les candidats concepts non valides.

Les résultats obtenus par le projet complet avec la base de concepts provenant d'AutoSlog sont comparables à ceux obtenus avec une base de concepts établie à la main par des experts. Par contre, AutoSlog a nécessité cinq heures de travail pour valider ses concepts, alors que l'auteur fournit une estimation de 1500 heures de travail pour constituer la base à la main. Le temps consacré à l'étiquetage du corpus d'apprentissage est précisé dans le travail présenté ci-dessous, 8 heures pour 160 textes, soit une semaine pour un corpus d'apprentissage suffisant.

Dans [Ril96a], E. Riloff décrit une méthode où il n'est plus nécessaire d'utiliser un corpus où des étiquettes "conceptuelles" telles que victime, etc. sont affectées au préalable aux groupes nominaux. Elle choisit, pour cela, d'extraire toutes les structures qui correspondent aux heuristiques utilisées précédemment, sans se préoccuper de savoir quelles sont les entités auxquelles elles se réfèrent. Ces structures sont ensuite classées suivant leur fréquence constatée dans le corpus. On obtient ainsi une liste très importante de candidats qu'il faut discriminer. La tâche est facilitée par le fait que les candidats sont ordonnés et que, très vite, il n'est plus très utile de chercher à retenir des candidats. Dans son expérience, l'auteur parcourt la liste ordonnée des candidats, ne retient que ceux qu'elle juge valides et s'arrête quand elle juge peu probable d'en rencontrer de nouveaux. Plus précisément, sur 11225 candidats initiaux, elle n'en conserve que 210 après avoir examiné les 1970 premiers, seuil où la fréquence des candidats retenus est trop faible pour qu'elle juge utile de continuer. Ce parcours manuel des candidats est également utilisé pour produire les patrons syntaxico-sémantiques. Les résultats obtenus avec cette méthode – l'outil s'appelle AutoSlog-TS – sont comparables à ceux d'AutoSlog. L'intérêt de ce travail est qu'il devient inutile d'étudier au préalable le domaine pour

48. Pas nécessairement un expert. Des expériences ont été réalisées avec des étudiants dans le rôle du juge (cf. [Ril96a]) ; les résultats obtenus sont acceptables, mais moins bons que dans le cas où le juge est un expert du domaine (cf. ibid.).

déterminer les groupes nominaux à examiner.

Dans [RS97], E. Riloff et J. Shepherd présentent une méthode pour construire de manière semi-automatique un lexique sémantique spécialisé sur les noms d'un domaine. L'objectif de ce lexique est de regrouper les termes qui sont liés à un concept très général du domaine. Les ensembles ainsi formés constituent des catégories lexicales associées aux concepts du domaine[49]. Pour ce faire, les auteurs définissent les concepts et leur associent quelques noms qui leur sont sémantiquement liés. Par exemple, pour le concept VEHICLE, elles utilisent les noms *airplane, car, jeep, plane* et *truck*. Ceci leur permet d'initialiser la construction d'un lexique, qui se définit plus comme un brouillon qui nécessite d'être travaillé pour être exploitable. Les catégories obtenues, bien que ne pouvant prétendre former des classes sémantiques, sont cependant tout à fait adaptées pour faciliter le travail préalable à l'utilisation de AutoSlog, à savoir l'étiquetage "conceptuel" des groupe nominaux du corpus qui représentent les objets de la tâche d'extraction d'informations.

De manière plus détaillée, leur démarche est la suivante :

1. pour chaque catégorie, déterminer une liste de cinq noms souches qui servent à initialiser le système ;

2. rechercher dans le corpus les phrases où ces noms sont présents. Soumettre ces phrases à une analyse lexicale, puis à une analyse syntaxique simple pour identifier les entités principales de la phrase (c'est-à-dire, les groupes nominaux, verbaux et prépositionnels) ;

3. considérer ensuite, pour chacun de ces noms souches, un voisinage constitué par le premier nom à gauche et le premier nom à droite. Les deux noms ainsi déterminés forment le contexte de la souche ;

4. grâce à une mesure de la fréquence relative des noms dans le contexte d'un nom souche par rapport à la fréquence de ce nom dans le corpus, déterminer quels sont les noms les plus fortement associés à la catégorie considérée ;

49. On parle ici de catégorie lexicale ou même de catégorie pour désigner de simples ensembles de noms liés à un même concept.

5. après avoir éliminé les noms trop peu fréquents dans le corpus, considérer les cinq noms les plus fortement associés à la catégorie étudiée. Recommencer le processus en considérant ces cinq nouveaux noms comme de nouvelles souches que l'on ajoute à la liste ;

6. après plusieurs itérations du processus, on obtient une liste de noms qui sont de bons candidats à apparaître dans un lexique sémantique au sein de la catégorie considérée.

Les groupes de noms ainsi constitués forment un ensemble comprenant des termes appartenant à la catégorie étudiée, comme **horse** pour ANIMAL, des termes ayant un lien avec des membres de la catégorie, par exemple **feather** ou **zoo** pour ANIMAL ainsi, bien sûr, que des termes n'ayant aucun rapport. Il est donc nécessaire de soumettre les listes de noms obtenues à un terminologue pour pouvoir les exploiter dans le cadre d'un lexique sémantique. De plus, les auteurs n'ont pas exploré les limites liées au choix des mots souches. Il serait intéressant de savoir en quoi leur choix influe sur la qualité du résultat obtenu.

Par ailleurs, les informations que l'on peut obtenir grâce aux travaux de E. Riloff précédemment exposés constituent également une aide précieuse pour la description d'une entrée lexicale : elles permettent de savoir, pour un lexème donné, les événements dans lesquels on peut le rencontrer.

Rajeev Agarwal [Aga95]

Au cours de son travail, R. Agarwal a cherché à établir une méthode semi-automatique pour construire une classification sur les noms et les verbes d'un domaine de vocabulaire particulier. Cette classification se donne pour but d'être une bonne approximation d'une classification sémantique[50] établie par un expert. Ici encore les connaissances lexicales nécessaires à la mise en œuvre de la méthode sont très réduites. Cette méthode se décompose en plusieurs étapes :

1. on commence par une analyse lexicale puis syntaxique du corpus, de façon à mettre en évidence des relations de dépendance syntaxique telles que sujet-verbe, verbe-complément ou les prépositions associées à un nom ;

50. Ici, au vu de la classification obtenue, il faudrait peut-être parler d'une ontologie du domaine.

2. ces relations sont utilisées comme attributs des termes étudiés ; par exemple pour un nom, on retient les deux verbes dont il est le plus fréquemment l'objet, les trois verbes dont il est le plus fréquemment le sujet et les deux prépositions avec lesquelles il est le plus fréquemment associé. Ceci permet de définir, pour les termes étudiés, un vecteur de « contextes » ;

3. une classification est ensuite effectuée pour regrouper les termes partageant des contextes similaires. On peut noter que la méthode utilisée permet aux termes sur lesquels se fait la classification d'appartenir à plusieurs classes ;

4. les classes ainsi formées sont validées en faisant appel à WordNet. On ne retient une classe que s'il existe dans WordNet un synset contenant l'ensemble des éléments de la classe construite précédemment.

Ces classes sont soumises au verdict d'un expert humain avant d'être considérées comme formant une classification sémantique sur les termes du domaine. La validation préalable des classes par WordNet allège le travail de l'expert, mais n'en change pas la nature. Les classes sémantiques construites pour les verbes et les noms permettent de typer ce que l'auteur appelle des « motifs lexico-sémantiques », ce qui signifie que sont identifiées des relations entre les classes sémantiques du domaine sous leurs formes syntaxiques. Ces relations (quelles classes de noms apparaissent en objet de quelles classes de verbes) peuvent servir à représenter des connaissances sur le domaine considéré. Ce dernier aspect n'est qu'effleuré à la fin du travail.

Cette étude montre qu'une approche simple du problème de l'acquisition lexicale peut donner des résultats corrects. Les classifications obtenues semblent pertinentes. On peut toutefois remarquer que l'auteur éprouve des difficultés à améliorer les performances de son système au cours des différentes expériences présentées.

3.2.2 Précision d'une classification

Le point commun des travaux que nous présentons dans cette section est de travailler à partir d'une classification sémantique et d'en préciser la structure pour la rendre plus adaptée à un domaine particulier. Pour ce faire, il est possible de chercher au sein d'une classe sémantique existante des ensembles

de mots qui en forment des sous-classes plus précises [VFP91] ou de tenter de combiner des classes existantes pour former une nouvelle classification [Bui97].

Paola Velardi et al. [VFP91]

L'objectif du travail présenté dans [VFP91] est, à partir d'une classification générale des mots présents dans les termes d'un domaine et d'une description des relations liant les membres de cette terminologie, de mettre en évidence une classification conceptuelle plus précise des éléments de ces termes. Par exemple, à partir d'une relation générale du domaine *appliquer(action, entité_animée)* et d'une instance de cette relation dans le corpus *appliquer(élevage, vache)*, les auteurs cherchent à obtenir la relation de niveau intermédiaire *appliquer(élevage, animal)*.

Velardi et al. définissent pour cela une classification hiérarchique des concepts généraux du domaine. Ils utilisent également un ensemble de relations du type $relgen_k(C_i, C_j)$, où C_i et C_j représentent des classes très générales du domaine. La démarche suivie est alors :

1. rechercher, pour un terme complexe du corpus, les relations générales qui peuvent expliciter ce qui lie ses éléments ;
2. répertorier ces relations instanciées par les constituants du terme comme éléments de la représentation la plus fine du domaine ;
3. après avoir généralisé les classes sémantiques des constituants du terme, répertorier ces mêmes triplets comme éléments de la représentation intermédiaire.

Cette méthode permet de déterminer de façon précise les couples de classes sémantiques liées par une relation à l'intérieur des termes complexes du domaine. La mise en œuvre d'une telle analyse sémantique des termes complexes est très simple. Le problème rencontré ici reste la nécessité d'une étude approfondie du domaine pour établir la hiérarchie initiale des concepts. Mais une fois celle-ci réalisée, le reste de la méthode est adaptable à tout domaine.

Paul Buitelaar [Bui97]

Dans cet article, P. Buitelaar plaide pour un lexique sémantique, le CoreLex, basé sur le formalisme du lexique génératif de Pustejovsky [Pus95]. Les entités représentées dans ce lexique sont des classes sémantiques polysémiques de noms. Une classe rassemble par exemple les noms qui représentent des animaux et de la nourriture. Ce lexique est destiné à servir de base à la construction d'un lexique spécifique à un domaine, par exemple en extrayant du corpus les relations qui permettront de préciser la représentation lexicale des noms du domaine.

Les classes sémantiques sont construites à l'aide des informations contenues dans la base lexicale WordNet. L'auteur considère les synsets les plus généraux du lexique comme autant de types sémantiques de base. Il recherche les combinaisons de ces types qui regroupent des noms polysémiques afin de former ses classes sémantiques polysémiques. Par exemple, il définit une classe `anm fod` qui contiendra les noms appartenant aux synsets ANIMAL et FOOD de WordNet. À ce stade, il obtient une liste de candidats à être considérés comme classes polysémiques. Il faut éliminer manuellement les classes formées non par des noms polysémiques, mais par des homonymes[51].

Une fois les classes polysémiques déterminées, l'auteur les dote d'une structure basée sur le lexique génératif, c'est-à-dire qu'il leur attribue une structure inspirée de la qualia du lexique génératif de Pustejovsky pour caractériser leurs attributs sémantiques, ainsi qu'un typage complexe basé sur les synsets utilisés précédemment. Ce typage, inspiré de celui utilisé par Pustejovsky, permet de représenter les différentes facettes sémiques d'un nom qui sont susceptibles d'être évoquées simultanément. Par exemple, dans la phrase : « le loup a mangé le mouton qui était dans le champ. », le nom **mouton** est associé aux deux concepts ANIMAL et FOOD. L'auteur reprend également les mécanismes d'héritage du lexique génératif. Il obtient ainsi un lexique qui fournit une base qui sera développée dans une application à un domaine particulier.

L'auteur se focalise ensuite sur cette adaptation d'un lexique à un domaine précis, c'est-à-dire sur la détermination des particularités du domaine que Co-

51. On rappelle qu'un nom est polysémique s'il représente des concepts différents mais qu'on peut relier. Par exemple un livre désigne aussi bien un objet que le contenu abstrait de cet objet. Par contre un homonyme représente des concepts totalement différents. Par exemple charleston désigne une ville ou une danse.

RELEX ne représente pas déjà. Pour ce faire, il cherche dans le corpus les structures syntaxiques qui sont associées aux éléments connus du lexique. Il détermine les mots inconnus qui sont liés aux mêmes structures dans le corpus. Ceci permet de déterminer à quelles classes des noms inconnus sont apparentés. On peut alors envisager de répéter l'opération pour améliorer encore le lexique. Il est également possible de trouver des prédicats spécifiques au domaine pour enrichir la représentation lexicale des classes ou des noms du domaine.

3.2.3 Relations ontologiques

Nous nous intéressons ici aux travaux qui cherchent à reconstruire une ontologie du vocabulaire d'un domaine particulier. Comme nous l'avons précisé en introduction, une ontologie est une représentation d'un domaine de connaissances. Nous voyons ici comment construire une telle entité, en mettant en évidence des ensembles de mots représentant des concepts similaires [Ass97, ZBHN97] ou en trouvant des relations prédicatives entre les éléments de ces ensembles [MPRT97]. Nous rappelons que nous considérons qu'une ontologie est une représentation des concepts d'un domaine ; pour une discussion plus approfondie sur les ontologies, on se référera à [Ras95, GG95, Bac96].

Houssem Assadi et al. [AB95, Ass96, Ass97]

Les travaux qui sont présentés ici ont pour objectif d'aider à la mise au point d'un système de consultation de documents techniques. Un tel système doit permettre à l'utilisateur de naviguer au sein de la documentation à l'aide d'un index ou des concepts du domaine. Pour cela, il faut disposer de la nomenclature des termes du domaine. Cet objectif est atteint par extraction de terminologie, ici à l'aide de Lexter [Bou94] (cf. 3.1.2, page 21). Une telle nomenclature est toutefois insuffisante pour exprimer les relations entre les différents termes, informations pourtant très utiles pour naviguer de façon efficace au sein de la documentation.

H. Assadi et al. proposent une méthode pour exploiter les informations contenues dans le réseau terminologique construit par Lexter afin d'établir une classification des adjectifs utilisés au sein des termes complexes, ainsi qu'une classification des noms qui y apparaissent associée à une description des concepts liés à ces noms.

La classification sur les adjectifs [AB95] s'effectue de façon très simple. On définit pour chacun d'eux un vecteur d'attributs qui décrit de quelles têtes l'adjectif étudié apparaît en expansion dans le réseau terminologique. On effectue ensuite une classification sur ces vecteurs. Ceci permet de mettre en évidence des adjectifs similaires.

Pour les noms [Ass96, Ass97], la classification procède de la même façon. On forme ainsi des classes de noms qui représentent les concepts du domaine. Cette première phase est toutefois suivie d'une analyse manuelle dite microscopique qui a pour objectif de décrire le concept représenté par la classe étudiée, à l'aide des termes où ses éléments apparaissent dans le réseau terminologique et des classes conceptuelles formées précédemment.

Benoît Habert et al. [ZBHN97, BHNZ97]

Dans ces travaux, les auteurs exposent leur travail visant à reconstruire une ontologie d'un domaine à partir d'une analyse sur corpus. Leur objectif est plus précisément de mettre en évidence des ensembles de termes qui forment une première esquisse d'une ontologie du domaine considéré[52]. Dans ce but, ils extraient du corpus les syntagmes nominaux à l'aide d'outils d'extraction de terminologie, ici Lexter [Bou94] (cf. 3.1.2, page 21) et AlethIP, puis ils utilisent le logiciel Zellig pour regrouper les mots qui partagent des contextes syntaxiques communs. Zellig construit un graphe dont les nœuds sont les mots étudiés, les arêtes qui relient ces nœuds représentant les contextes syntaxiques que les mots possèdent en commun. Les arêtes sont étiquetées par ces contextes. Zellig ne conserve par la suite que les arêtes significatives, c'est-à-dire celles qui représentent un certain nombre de contextes. On obtient ainsi des ensembles de mots qui sont représentés sous la forme de composantes connexes dans le graphe, voire de cliques. Ces ensembles sont des candidats à former des classes sémantiques. On peut de plus, si une partie des mots appartient déjà à une catégorie sémantique connue, étendre ces catégories à de nouveaux mots, par voisinage dans le graphe. Les associations entre les termes ainsi déterminées semblent pertinentes quand on les compare aux informations contenues dans une ontologie construite par des experts du domaine. Les nouveaux termes paraissent également être généralement affectés à de bonnes classes sémantiques.

52. Ici, le vocabulaire médical.

Cette évaluation est faite dans [BHNZ97] par rapport à deux ontologies existantes du même domaine.

Yannick Toussaint et al. [MPRT97]

Dans [MPRT97], les auteurs présentent une méthode d'extraction de connaissances à partir de textes. Ici, les textes sont des résumés d'articles traitant de la culture et du conditionnement des fruits.

Les auteurs utilisent le logiciel FASTR [Jac97] (cf. 3.1.3, page 22) pour extraire du corpus les éléments de la terminologie du domaine. En fait, Toussaint et al. utilisent un thesaurus du domaine, AGROVOC, pour déterminer les formes canoniques des termes et cherchent, à l'aide de FASTR, les textes où ils apparaissent sous cette forme ou sous une autre. Après validation experte des candidats termes extraits, les auteurs construisent des *clusters* de termes. Ces ensembles sont obtenus par une classification ascendante effectuée à l'aide d'un coefficient d'association entre deux termes, basé sur leurs cooccurrences dans un même texte. Chacun de ces *clusters* forme un graphe où les arêtes déterminent les liens au sein du cluster, appelés liens internes. Ces arêtes sont pondérées par le coefficient d'association calculé précédemment. Les auteurs gardent également la trace des liens les plus forts qui lient les éléments d'un cluster avec ceux d'autres clusters, appelés liens externes. La taille maximale des clusters est limitée par l'opérateur.

Les ensembles de termes ainsi constitués sont ensuite étudiés par un expert du domaine. À la suite de cette expertise, il est possible de structurer les éléments des clusters à l'aide de relations classiques de sémantique lexicale (cf. introduction) et de relations prédicatives (cf. 3.4, page 48). Les relations associées aux liens, internes ou externes, d'un cluster ne sont donc pas toutes du même type. Il est ainsi possible de trouver des synonymies, des hyperonymies, mais également différents prédicats pour un même cluster, au contraire des travaux de Assadi et al. et Habert et al. où les ensembles obtenus sont plus homogènes. Par contre, la représentation obtenue par Toussaint et al. est plus exhaustive, grâce à ces relations prédicatives mises en évidence lors de l'analyse experte des clusters. Toutefois, cette dernière méthode nécessite de disposer d'un thesaurus des termes du domaine, ce qui nuit à portabilité de la méthode et la rend dépendante de la qualité du thesaurus utilisé ; les auteurs présentent d'ailleurs quelques problèmes rencontrés, liés aux insuffisances du thesaurus.

3.2.4 Relations linguistiques

L'objectif des études décrites dans cette section est de mettre en évidence des classes sémantiques qui sont associées à un concept mis en jeu lors de relations linguistiques connues. On identifie une relation syntaxique et on cherche à mettre en évidence quelles sont les classes sémantiques qui peuvent être associées par cette relation [Res93, BPV92]. On peut également se focaliser sur une relation sémantique et rechercher les relations syntaxiques qui lui sont associées dans le corpus [Mor97].

Philip Resnik [Res93]

Pour exprimer des liens entre des classes sémantiques, P. Resnik a mis au point la mesure d'*association sélectionnelle*. Si, dans les phrases, les prédicats établissent des contraintes sémantiques sur leurs arguments, il est par contre difficile d'établir, pour un argument donné, quel est le type sémantique effectivement manipulé dans une phrase donnée.

Pour pouvoir déterminer la classe sémantique des arguments d'un prédicat, la méthode couramment employée est d'observer quelles sont les instances de ces arguments les plus représentatives, par exemple par un calcul d'*information mutuelle*, et ensuite d'en inférer la classe sémantique des arguments. L'*association sélectionnelle* a elle pour objet de calculer, pour un verbe et un argument donnés, la classe sémantique de l'argument considérée dans cette relation. L'utilisation d'un tel outil ne peut se faire qu'avec l'aide d'une classification sémantique complète et pertinente. C'est pourquoi Resnik utilise la taxinomie WordNet sur les noms. Il peut ainsi déterminer dans une structure argumentale quelle est la classe sémantique la plus pertinente pour l'argument.

Dans son étude, Resnik établit, par exemple, la classe la plus pertinente pour *baseball*, sachant le prédicat utilisé, dans les phrases suivantes :

1. I hit a baseball to John and he caught it.
2. I played some baseball yesterday afternoon.
3. I watched some baseball yesterday afternoon.

L'utilisation de l'*association sélectionnelle* permet de déterminer que la classe sémantique la plus probable est respectivement :

1. OBJECT
2. GAME
3. DIVERSION

L'*association sélectionnelle* entre un prédicat p_i et une classe sémantique c s'exprime de la façon suivante :

$$A(p_i, c) = \frac{1}{n_i} p(c|p_i) \log \frac{p(c|p_i)}{p(c)}, \text{ où } n_i \text{ s'exprime ainsi } n_i = \sum_c p(c|p_i) \log \frac{p(c|p_i)}{p(c)}.$$

Dans cette formule, $p(c|p_i)$ est la probabilité qu'un élément de la classe c apparaisse comme argument du prédicat p_i, $p(c)$ est la probabilité qu'un élément de la classe c apparaisse comme argument d'un prédicat quelconque et n_i représente un coefficient de normalisation qui permet, pour chaque prédicat p_i, que la somme des associations de p_i avec toutes les classes sémantiques du lexique soit égale à 1 ($\sum_c A(p_i, c) = 1$).

Pour illustrer les utilisations possibles de l'association sélectionnelle, Resnik en fait également usage pour parenthéser les noms composés. Un problème posé par les noms composés comprenant plus de deux noms est que l'on ne sait pas retrouver la façon dont ils sont formés. Ainsi, quand on rencontre un nom composé de la forme N1 N2 N3, doit-on l'interpréter comme [N1 N2] N3 ou comme N1 [N2 N3] ? Le premier cas se rencontre plus souvent que le second, mais la proportion étant de deux tiers-un tiers, cette structure reste très ambiguë.

Pour résoudre ce problème, Resnik propose d'étudier la force de l'association sélectionnelle entre les éléments des deux couples [N1 N2] et [N2 N3][53]. Les résultats obtenus sur ce point sont intéressants, mais encore insuffisants. En effet, il existe des cas où le critère ne permet pas de déterminer la solution.

53. Faute de disposer d'un véritable prédicat, l'auteur choisit de calculer l'association sélectionnelle entre les deux éléments d'un composé [N1 N2] en assimilant N2 au prédicat dans son calcul.

Toutefois on peut rappeler que, même pour un expert, il est souvent difficile de donner un résultat sans informations sur le contexte d'utilisation du composé ; aussi doit-on considérer ces premiers travaux comme encourageants. Il est intéressant de noter que l'une des limites de cette approche du parenthésage des composés, ainsi que Resnik le souligne, est due au fait que la relation entre les constituants des composés n'est pas identifiée.

Paola Velardi et al. [BPV92]

Dans [BPV92], P. Velardi et al. définissent une méthode permettant, à partir d'un corpus spécialisé et pour une relation syntaxique donnée (par exemple N_1 prep N_2), de calculer les classes conceptuelles[54] les plus susceptibles de se retrouver dans cette relation. Les résultats ainsi obtenus sont appliqués pour aider à résoudre l'interprétation de certaines structures syntaxiquement ambiguës.

Pour ce faire, les auteurs définissent une classification conceptuelle des termes simples de leur corpus en n classes sémantiques $C_1..C_n$. Ils recensent de plus, pour certaines relations syntaxiques élémentaires et d'arité deux, notées $syntrel_i(x, y)$, l'ensemble des occurrences de ces dernières dans le corpus. Ils calculent ensuite, pour une des relations syntaxiques données, un tableau contenant, pour chaque couple de classes (C_k, C_l), la proportion de couples (T_1, T_2) tels que :

$$T_1 \in C_k \text{ et } T_2 \in C_l \text{ et } syntrel_i(T_1, T_2).$$

Le résultat ainsi obtenu est une estimation de la probabilité conditionnelle $P(C_k, C_l|syntrel_i)$, et a pour valeur :

$$\frac{freq(C_k, synrel_i, C_l)}{freq(syntrel_i)}.$$

Ce calcul permet de déterminer, pour une relation syntaxique donnée, la proportion des termes réalisant cette relation syntaxique, appartenant à un couple de classes donné.

54. Dans les travaux de P. Velardi et al., les classes conceptuelles correspondent à des classes sémantiques.

Ces résultats peuvent être utilisés pour résoudre des problèmes d'ambiguïté syntaxique. Lors d'une analyse syntaxique, certaines structures sont difficiles à interpréter parce qu'il existe plusieurs solutions possibles et que les informations grammaticales dont on dispose sont insuffisantes. Ainsi, dans le cas d'une structure du type $VN_1 prep N_2$ telle que « *sottomettere la richiesta all autorita competente* »[55], il est impossible de distinguer, à l'aide de critères syntaxiques, si la préposition se réfère au couple (N_1, N_2) ou au couple (V, N_2). C'est ici que la force du lien sémantique permet de déterminer de ces deux interprétations laquelle est la plus plausible. Les résultats de cette méthode peuvent cependant conserver tout ou partie de l'ambiguïté de la structure syntaxique initiale, mais lorsque cette ambiguïté est levée, les résultats sont satisfaisants.

Cette méthode de calcul d'associations préférentielles à partir d'un corpus permet donc de déterminer les couples les plus souvent utilisés parmi les classes sémantiques possibles des arguments d'une relation syntaxique donnée. On peut ensuite, pour une relation donnée et une classe sémantique d'un de ses arguments, exprimer une préférence sur celle du second argument. La classification utilisée reste très simple : P. Velardi et al. ont utilisé une dizaine de classes pour leurs expériences.

Emmanuel Morin [Mor97]

Dans cet article, l'auteur présente une méthode pour rechercher des relations sémantiques au sein d'un corpus. Son objectif est de retrouver ces relations à partir de schémas syntaxiques qui en sont des indicateurs.

Pour ce faire, il illustre son propos par la description de la démarche suivie pour détecter des hyponymies sur corpus. Il est nécessaire à la mise en œuvre de sa méthode de disposer d'un ensemble de couples de mots liés par une telle relation sémantique, ensemble qui sert à amorcer le système.

La démarche suivie consiste à rechercher dans le corpus les contextes syntaxiques où les deux termes de tels couples apparaissent. L'ensemble de contextes ainsi constitué est réduit à un ensemble de patrons syntaxiques candidats à être des indicateurs de la relation sémantique recherchée. Ces patrons sont ensuite validés par un terminologue. On les utilise alors pour rechercher d'autres couples de mots qui apparaissent dans des contextes qu'ils couvrent. Ces nou-

55. « soumettre la requête à l'autorité compétente ».

veaux couples sont également validés par le terminologue. Puis le nouvel ensemble de couples de termes ainsi défini est utilisé pour rechercher de nouveaux patrons syntaxiques. Le processus général se repète jusqu'à ce qu'on ne trouve plus de nouveaux patrons syntaxiques. On obtient au final un ensemble de couples de mots liés par la relation sémantique étudiée.

3.3 Mise-à-jour du lexique

Mettre à jour un lexique, c'est ajouter, enlever ou modifier des entrées lexicales. Le choix de ces entrées est lié à l'usage prévu du lexique. Dans le cas, fréquent, où on cherche à l'adapter au vocabulaire d'un langage donné, ce sont les comportements constatés de ce langage qui déterminent les changements à effectuer. On regarde si le comportement des mots dans le langage est en contradiction avec ce qui est prédit par le lexique et, si oui, on modifie ce dernier afin qu'il autorise ces nouveaux comportements langagiers. Par exemple, si on dispose d'un lexique qui ne référence que le sens CHEVAL pour le mot **horse** et qu'on se place dans l'étude du langage utilisé dans les romans policiers, le sens marginal de **horse**, HÉROÏNE, pourra être utilisé dans les textes et devra donc être intégré au lexique.

Nous montrons comment il est possible d'associer à un mot un sens préférentiel parmi l'ensemble de ses sens possibles. Ces préférences sont déterminées à l'aide d'autres mots que l'on suppose apparentés [Res95a, Res95b] ou en s'appuyant sur des relations sémantiques connues [Pic96, Séb96]. Nous voyons ensuite comment enrichir un lexique par ajout de nouvelles entrées lexicales [PAB93].

3.3.1 Désambiguïsation

Désambiguer un mot ambigu, tel **bank** en anglais, consiste à chercher, parmi ses divers sens, celui qui est utilisé dans une manifestation concrète telle qu'une phrase. La désambiguïsation suppose donc une représentation des différents sens des mots étudiés et un *contexte* qui permette de trouver quel sens du mot concerné est privilégié. Le terme contexte signifie ici tout ce qui est extérieur au mot considéré. Les travaux suivants s'intéressent à ce problème.

Philip Resnik [Res95a, Res95b]

Dans ces travaux, P. Resnik étudie les groupes de noms tels que ceux rencontrés au sein d'un thesaurus. Ces ensembles associent des noms ayant quelque chose en commun, sans que la ou les relations qui les réunissent soient explicitées. Ainsi par exemple, les mots **doctor** et **nurse** désignent tous deux des gens travaillant dans le domaine médical, mais également respectivement un titulaire d'un doctorat et une garde d'enfant. Si on recherche pourquoi ces deux noms sont regroupés au sein d'une même catégorie, il est naturel de penser que c'est parce qu'ils désignent tous deux des professionnels de la médecine. C'est cet objectif que Resnik se propose d'étudier.

La première étape de ce travail [Res95a] consiste à définir une méthode pour déterminer, parmi les concepts que peuvent représenter deux noms, ceux qui représentent le mieux ce qui les rassemble. Ainsi pour **doctor** et **nurse**, ce sera le fait qu'ils désignent tous deux des professionnels de santé, mais pour **doctor** et **lawyer**, ce sera le fait que ces deux noms désignent des professionnels en général. La méthode est ensuite généralisée aux groupes de noms [Res95b], pour déterminer, dans un ensemble d'éléments ayant une parenté sémantique, le sens de chacun d'eux effectivement considéré dans le regroupement. Nous détaillons successivement ces deux travaux.

Dans [Res95a], Resnik cherche à évaluer la similarité entre deux noms au sein d'une taxinomie et utilise la base lexicale WordNet comme référence. C'est la taxinomie définie par la relation IS-A qui est utilisée. Pour déterminer parmi les synsets de WordNet celui qui représente le mieux ce que les deux noms ont en commun, l'auteur compare les trois méthodes suivantes :

1. on compte les arêtes de l'arbre des synsets. On considère ici que deux concepts représentés par les synsets de WordNet sont d'autant plus similaires qu'il existe dans l'arbre de la taxinomie un chemin court qui les relie. Le problème lié à cette approche est que, dans la taxinomie de WordNet, les arêtes qui relient deux synsets n'expriment pas une distance sémantique homogène ;

2. on considère que chaque synset a une probabilité donnée d'apparaître au sein du corpus[56]. On recherche alors, parmi les synsets communs

56. La probabilité est estimée à partir d'une analyse sur corpus.

aux noms considérés, celui qui est le plus probable. Cette mesure de la similarité a pour principal inconvénient de sélectionner des synsets trop généraux ;

3. la méthode, proposée par l'auteur, est basée sur l'information. On recherche le synset qui représente le plus d'information sur les deux termes considérés. Pour ce faire, Resnik utilise une mesure de l'information portée par chaque synset de WordNet, mesure qui croît quand on descend dans l'arbre du lexique, puisque les concepts représentés par ces synsets sont plus spécifiques.

La mesure proposée par l'auteur permet de saisir des relations intéressantes sur les noms étudiés. Le seul problème est lié à WordNet qui référence des sens peu communs pour certains noms (exemple : DRUG pour **horse**).

Comme nous l'avons dit précédemment, la méthode est étendue aux groupes de noms dans [Res95b]. Resnik cherche à déterminer, pour un ensemble de noms regroupés parce qu'ayant une « parenté » sémantique, le sens de chacun de ces noms qui est effectivement considéré. De tels renseignements peuvent, par exemple, permettre de mettre à jour une taxinomie préexistante sur les noms à l'aide d'une classification nouvellement déterminée.

La méthode utilisée peut se résumer ainsi :

1. pour chaque nom, rechercher quels sont les synsets de WordNet qui expriment le mieux la similarité avec chacun des autres noms considérés selon la méthode décrite ci-dessus ;

2. pour chacun des sens du nom étudié, chaque fois qu'un synset qui représente un de ses parents dans l'arborescence de WordNet est sélectionné, il contribue à rendre ce sens plus « crédible ».

On peut ainsi, pour chaque nom, savoir lequel de ses sens dans WordNet peut être considéré comme le plus « proche » des autres noms. De plus, à chacune de ses interprétations est associé un coefficient qui permet de juger de sa crédibilité. Les résultats obtenus par cette méthode sont illustrés par de nombreux exemples. Les groupes de noms étudiés sont issus de diverses sources, et sont essentiellement obtenus par des travaux de regroupements de mots, tels

que ceux présentés en 3.2.1 (page 24). Une évaluation plus formelle est également présentée où les résultats fournis par la machine semblent satisfaisants.

Une telle méthode est essentiellement intéressante quand il s'agit de déterminer quel sens d'un mot est utilisé. Elle nécessite toutefois de savoir l'ensemble des sens possibles du mot étudié et de connaître d'autres mots utilisés dans des contextes similaires. Elle peut notamment s'appliquer dans le cas où on travaille sur un ensemble de mots extraits automatiquement d'un corpus par regroupement. Il est possible d'envisager son utilisation pour adapter une base lexicale de l'ampleur de WordNet à un domaine précis, c'est-à-dire ne conserver que les synsets pertinents pour un domaine particulier. Dans cette dernière tâche, cette méthode peut être envisagée en assistance de l'expert chargé de l'adaptation du lexique.

Ronan Pichon et Pascale Sébillot [Pic96, Séb96]

Ces travaux se situent dans la continuité de ceux effectués sur l'interprétation de séquences binominales complexes par C. Fabre et P. Sébillot [Fab96] (cf. 2.3, page 16). Le modèle élaboré permet de calculer, pour un composé nominal quelconque, l'ensemble de ses interprétations possibles, hors domaine et sans prendre en compte d'informations sur le contexte[57] au sein duquel il est exprimé. Ce qui est décrit ici est une méthode pour discriminer, parmi les interprétations fournies par le modèle général, celles qui sont les plus vraisemblables dans un domaine donné.

Pour ce faire, on se donne un corpus de noms composés du domaine, ainsi qu'un lexique sémantique de référence, ici WordNet. On affecte à l'aide du corpus un coefficient d'association aux couples de classes sémantiques que peuvent former les composés du domaine. Par exemple, au composé *sand bank*, on peut associer, entre autres, les couples : (FORTITUDE, INSTITUTION), (SOIL, SLOPE), etc. On observe ensuite les interprétations fournies par le modèle général pour chaque composé nominal. À chacune d'elle correspond un couple de classes sémantiques, et on discrimine les interprétations suivant le coefficient d'association calculé pour ces couples. Le filtre est détaillé dans [Pic96]

57. Contexte au sens large ; cela désigne aussi bien le contexte linguistique, la phrase ou le texte où ces composés peuvent apparaître que le domaine de langage dont le composé fait partie.

et son intégration dans le modèle d'interprétation des composés est discuté dans [Séb96].

3.3.2 Acquisition de nouvelles entrées lexicales

Acquérir une entrée lexicale, c'est déterminer une association entre un lexème et un concept (cf. introduction). Les travaux que nous présentons ici ne s'intéressent qu'au cas où le lexème est inconnu dans le lexique, c'est-à-dire qu'on ne cherche pas à associer de nouveaux concepts à un lexème déjà référencé dans le lexique.

James Pustejovsky et al. [PAB93]

Nous avons déjà mentionné les travaux de J. Pustejovsky [Pus95] sur la représentation lexicale (cf. 2.2, page 13). Dans ces travaux, l'auteur utilise un treillis pour représenter les relations entre les entrées lexicales. Pour les noms, ce treillis est notamment construit sur la base d'une relation « sorte de » entre les noms. C'est cette seule relation qui est considérée dans le travail que nous présentons ici, où les auteurs proposent une méthode pour acquérir la description lexicale de nouveaux noms.

La première information que l'on recherche, pour un nom N inconnu, est sa place dans la taxinomie du lexique. Pour cela, les auteurs émettent l'hypothèse suivante : l'une des relations reliant les constituants d'un nom composé est la relation « sorte de », où le modifieur est une instance de la classe du nom tête (par exemple *C language*). Si un mot non répertorié doit être attaché au lexique, son père dans la taxinomie sera probablement un nom tête d'un composé dont le nom inconnu est le modifieur. On forme ainsi un ensemble de noms candidats à être un ancêtre de N dans la taxinomie du lexique.

Par ailleurs, on se base sur le fait que deux noms liés par une relation « sorte de » sont proches sémantiquement et apparaissent en arguments des mêmes prédicats. On calcule donc l'*information mutuelle* entre N et les verbes le prenant comme complément d'objet, et on fait de même avec les noms candidats et ces mêmes verbes, l'*information mutuelle* permettant de déterminer avec quelle force deux termes sont liés. Une fois ces calculs effectués, pour chaque couple N/nom candidat, on effectue la somme des produits des informations

mutuelles entre N et un verbe et entre le candidat et le même verbe. On obtient donc la formule suivante, où N est le nom inconnu, N_j le nom candidat, $I(N;V)$ l'information mutuelle entre N et V et V_i décrit l'ensemble des verbes prenant N en complément d'objet.

$$S(N_j) = \sum_{V_i} I(N;V_i)I(N_j;V_i).$$

Le candidat obtenant le total $S(N_j)$ le plus important est alors sélectionné comme père du nom inconnu dans la taxinomie du lexique.

Le nom hérite de certains prédicats liés aux attributs sémantiques de son père. La structure est complétée par une sélection parmi les prédicats prenant le plus couramment l'inconnu en argument.

3.4 Prédicats d'un domaine

Dans cette partie, nous présentons des travaux dont l'objet est l'étude des *prédicats* d'un domaine à partir d'une analyse sur corpus. On considère généralement que le sens d'une structure syntaxiquement complexe peut s'exprimer à l'aide d'un prédicat [Jac90]. Celui-ci détermine les relations entre les groupes de mots de cette structure. Ce rôle est généralement tenu par un verbe et le prédicat est doté d'une *structure argumentale*. Cette dernière est une liste d'arguments, auxquels on associe habituellement un rôle dit thématique (tel que agent, thème, etc.) qui explicite le rôle de l'argument considéré dans la relation exprimée par le prédicat. De plus, chacun de ces arguments doit obéir à certaines contraintes sémantiques. Par exemple, la phrase «le chat est noir» peut s'exprimer de la façon suivante : A POUR COULEUR(chat, noir) ; ici, noir doit désigner une couleur. Les informations qui permettent de juger du type sémantique des mots du langage sont regroupées dans le lexique. Aux entrées correspondant aux prédicats, généralement des verbes, on associe les contraintes qu'ils exercent sur leurs structures argumentales. Aux lexèmes qui jouent le rôle d'arguments, généralement des noms, on associe les contraintes sémantiques qu'ils satisfont.

De nombreux travaux s'intéressent à l'étude de la structure argumentale des prédicats. L'objectif est de retrouver, à partir de phrases où un prédicat étudié est présent, les structures argumentales qui lui sont associées. Nous

présentons tout d'abord des travaux qui cherchent uniquement à mettre en évidence la forme syntaxique de ces structures argumentales [Man93, BC97], par exemple V NP PP, qui décrit que le verbe prend un complément d'objet et un groupe prépositionnel en argument. Nous décrivons ensuite comment il est possible de trouver les informations sémantiques associées aux arguments [PS96] ou leurs rôles thématiques [PSDM94], ce qui, comme nous le verrons, nécessite un travail préalable conséquent. Nous terminons par la question des verbes supports de nominalisations, c'est-à-dire des verbes associés à un nom déverbal[58] qui tient un rôle prédicatif (lié au verbe dont il dérive), par exemple *faire* dans : proposer/faire une proposition. De tels travaux permettent d'identifier quand un nom déverbal a un rôle de substantif et quand il a un rôle prédicatif [GT95, Dra95].

3.4.1 Acquisition de sous-catégorisations

Nous nous intéressons ici à des travaux dont l'objectif est d'acquérir sur corpus la structure de sous-catégorisation de prédicats (essentiellement des verbes). La sous-catégorisation est la forme syntaxique que prend la structure argumentale des prédicats étudiés, information qui peut être intégrée dans un lexique.

Christopher Manning [Man93]

Le travail présenté ici a pour but d'affecter automatiquement aux verbes la forme syntaxique de leurs structures argumentales. C. Manning part du constat qu'il est très long et très difficile de construire à la main un dictionnaire qui recense les structures de sous-catégorisation associées aux verbes du lexique. Il suppose qu'une analyse simple sur corpus peut apporter des informations suffisantes pour construire automatiquement un tel dictionnaire. Sa démarche est très simple :

1. il énumère les classes de structures de sous-catégorisation possibles. Il en utilise 19, qu'il suppose suffisantes pour couvrir l'ensemble des possibilités ;

58. Nom qui dérive morphologiquement d'un verbe. Le phénomène est appelé nominalisation.

2. après avoir affecté à chaque élément du corpus sa catégorie grammaticale, il utilise un analyseur syntaxique volontairement rudimentaire pour affecter à chaque occurrence des verbes du corpus une classe de structures de sous-catégorisation ;

3. il utilise ensuite un filtre statistique basé sur la loi binomiale pour ne conserver que les structures statistiquement pertinentes pour chaque verbe.

Les résultats obtenus sont ensuite comparés aux données présentes dans un dictionnaire, l'OALD[59]. Le bilan semble relativement satisfaisant (90% de précision[60], 43% de rappel[61]) au vu de la méthode utilisée et du fait que toutes les structures référencées dans le dictionnaire ne sont pas représentées dans le corpus, et ne peuvent donc y être apprises.

Ces résultats servent donc à déterminer les structures de sous-catégorisation associées aux verbes du corpus, afin de les stocker dans un lexique. De telles informations permettent notamment de faciliter une analyse des dépendances prédicats-arguments dans les phrases.

Ted Briscoe et John Carroll [BC94, BC97]

Dans leurs travaux, T. Briscoe et J. Carroll étudient également la possibilité de reconstruire la forme syntaxique de la structure argumentale de verbes à partir de corpus. La différence est que l'ensemble des formes possibles de ces structures argumentales est beaucoup plus riche que dans les travaux précédents [Man93][62]. De même, l'analyseur syntaxique utilisé est plus sophistiqué puisque le résultat de ce travail rassemble un nombre bien plus important de structures argumentales possibles. Cependant, la démarche générale reste très proche ; notamment, le même filtre statistique est utilisé pour discriminer les résultats.

59. Oxford Advanced Learner's Dictionary of Current English.

60. La précision est la proportion de bonnes réponses fournies par le système par rapport à l'ensemble des réponses fournies.

61. Le rappel est la proportion de bonnes réponses fournies par le système par rapport aux bonnes réponses attendues (i.e. connues).

62. Les auteurs utilisent 140 schémas extraits de dictionnaires, à comparer aux 19 classes utilisées par Manning.

3.4.2 Acquisition de structures argumentales

Les travaux que nous présentons ici se donnent pour objectif de mettre en évidence des informations sur le type sémantique [PS96] ou le rôle thématique [PSDM94] des éléments de la structure argumentale de prédicats.

Victor Poznański et Antonio Sanfilippo [PS96]
Dans ce travail, les auteurs présentent une méthode pour associer à un verbe prédicatif les types sémantiques des arguments auxquels il est lié dans un corpus. En fait, il s'agit, pour chaque verbe polysémique, de déterminer quel est le sens qui est utilisé dans les phrases du corpus à l'aide de sa structure argumentale. Les prérequis de leurs travaux sont les suivants :

- un corpus dont les phrases sont parenthésées, c'est-à-dire où les dépendances syntaxiques sont connues. Le choix des auteurs s'est porté sur le *Penn TreeBank* ;

- un ensemble de catégories sémantiques à associer aux mots du corpus. Pour ce faire, les auteurs utilisent le *Longman Dictionary of Contemporary English*[63], dictionnaire organisé en un arbre dont les nœuds représentent des classes sémantiques[64]. Ce dictionnaire leur permet de disposer en outre des schémas de sous-catégorisation des verbes du dictionnaire.

À partir de ces données, le travail présenté permet aux auteurs de passer des relations syntaxiques de la structure argumentale, par exemple ((FACES VBZ) (NP (CHARGES NNS)))[65], aux types sémantiques associés aux éléments de cette relation. Le typage sémantique se fait à des niveaux différents suivant la catégorie du mot étudié. Pour les verbes, les auteurs utilisent l'ensemble des

63. Abrégé en LDOCE

64. Les concepts associés à ces classes sémantiques sont de plus en plus précis à mesure qu'on descend dans l'arbre. Il y a quatorze nœuds en haut de l'arbre et plus de mille deux cents feuilles.

65. Ceci est un extrait du *Penn Treebank* dans lequel le verbe FACES prend comme complément d'objet le groupe nominal formé par le nom CHARGES. L'étiquette VBZ indique que FACES est un verbe, NNS que CHARGES est un nom au pluriel et NP que CHARGES forme un groupe nominal.

types sémantiques disponibles dans le lexique ; ils descendent donc jusqu'aux feuilles de l'arbre formé par le lexique. Pour les arguments, c'est-à-dire les noms, ils ne considèrent que les catégories sémantiques les plus générales (i.e. les nœuds directement attachés à la racine de l'arbre).

Plus précisément, à chaque occurrence d'un prédicat dans le corpus, on associe sa structure argumentale constatée et on cherche, au sein du LDOCE, le (ou les) sens du verbe qui possède(nt) un schéma de sous-catégorisation correspondant. Ceci permet, d'une part, de choisir un ou plusieurs sens plausible(s) pour le verbe et, d'autre part, d'associer un typage sémantique au schéma de sous-catégorisation concerné.

Au final, on dispose, pour chaque verbe du dictionnaire, non seulement de la forme syntaxique de sa structure argumentale, mais également d'informations sur la catégorie sémantique de ses arguments.

Patrick Saint-Dizier et al. [PSDM94]

Dans cet article, les auteurs décrivent une méthode pour trouver, dans un corpus de textes, des informations sur les prédicats du domaine. L'objectif est d'identifier les termes importants du domaine étudié (i.e. des noms), les états ou actions qui leur sont associés (i.e. des prédicats) et le rôle que jouent les termes dans ces états ou actions (i.e. le rôle thématique qui leur est associé).

Pour ce faire, les auteurs se livrent à une analyse approfondie des besoins de leur système. Tout d'abord, ils établissent une représentation des connaissances linguistiques nécessaires :

1. une représentation de la sémantique des prédicats basée sur le travail de R. Jackendoff [Jac90] et s'appuyant sur la classification de B. Levin [Lev93],

2. une classification des concepts du domaine basée sur les rôles thématiques qui leur sont associés.

De plus, Saint-Dizier et al. s'intéressent à la topologie des textes qu'ils ont à étudier, de courtes description de projets, pour mettre en évidence les marqueurs qui indiquent des articulations dans ces textes, de façon à savoir le type d'information apportée par un segment du texte. Ils identifient également des marqueurs qui indiquent typiquement des relations thématiques entre un

prédicat et un argument. Ces marqueurs sont généralement des mots-clés ou des syntagmes qui jouent ce rôle.

Les auteurs se livrent ensuite à une analyse syntaxique des textes étudiés pour retrouver les relations prédicatives qui y sont exprimées. Ils recherchent donc quels sont les prédicats utilisés dans les phrases du texte. À ces prédicats sont associés leurs arguments et leurs rôles thématiques en fonction des connaissances linguistiques préalables ou des relations mises en évidence par les articulations du texte.

3.4.3 Verbes supports de nominalisations

Certains verbes possèdent une forme nominale et celle-ci est utilisée parfois comme un nom commun, parfois comme une forme nominalisée du verbe. Dans ce dernier cas, le nom déverbal est associé dans la phrase à un verbe spécifique, dit verbe support (par exemple *to propose = to make a proposal*). Les travaux qui suivent s'attachent à déterminer ces verbes supports. Bien que cela ne soit pas à proprement parler des travaux d'acquisition d'informations sur les prédicats d'un domaine, ils peuvent néanmoins y mener. En effet, si un nom déverbal est employé avec son verbe support, c'est alors ce nom qui joue, d'un point de vue sémantique, le rôle du verbe dont il dérive, notamment le rôle de prédicat de la phrase. C'est pourquoi il est utile de préciser dans le lexique à quel verbe support un nom déverbal est lié, puisqu'il permet d'accéder à un sens possible du nom.

Gregory Grefenstette et Simone Teufel [GT95]

Dans leurs travaux, G. Grefenstette et S. Teufel [GT95] étudient le processus de nominalisation des verbes (c'est-à-dire la formation de noms déverbaux). Ils proposent une méthode permettant de mettre en évidence de façon automatique les verbes supports des nominalisations. En effet, si le processus de nominalisation est un phénomène langagier courant, cette opération ne fait pas appel à un verbe unique : chaque nom déverbal est utilisé avec le verbe support qui lui est propre.

Ils émettent l'hypothèse suivante : si, dans une phrase où apparaît un déverbal, le verbe principal le prend pour complément d'objet [66], il est possible que ce verbe principal soit le verbe support recherché. Par conséquent, si on recense tous les verbes répondant à cette condition pour un déverbal donné, le verbe support sera certainement l'un de ceux les plus fréquemment rencontrés.

Après une première expérience, il s'avère que cette hypothèse n'est pas suffisante. En effet, si le verbe support recherché figure parmi les candidats les plus fréquemment rencontrés, il est par contre rare qu'il soit le premier d'entre eux. Pour renforcer leur hypothèse et différencier l'utilisation d'un déverbal en tant que forme nominalisée d'un verbe de celle où il n'est considéré que comme un nom commun, Grefenstette et Teufel utilisent le fait que le déverbal hérite de la structure argumentale du verbe dont il dérive. Ils cherchent alors les prépositions les plus souvent rencontrées à la fois dans les phrases contenant le déverbal et dans celles contenant le verbe dont il dérive. Ils effectuent ensuite l'opération précédente en ne considérant que les verbes qui ont dans leurs structures de sous-catégorisation les prépositions reconnues comme significatives.

Les résultats obtenus dans ce cas sont pertinents. En effet, les verbes extraits sont fréquemment les verbes supports recherchés.

Mark Dras [Dra95]

Dans cet article, M. Dras approfondit les travaux de Grefenstette et Teufel présentés précédemment. Lui aussi se donne pour objectif de déterminer de façon automatique les verbes supports utilisés au cours de chaque processus de nominalisation.

Son hypothèse complète celle utilisée dans [GT95], car il considère que le verbe support recherché fait partie des verbes les plus fréquemment utilisés avec le nom déverbal étudié, mais est également un des verbes les plus fréquemment rencontrés avec l'ensemble des déverbaux du corpus. Ceci a pour objectif de résoudre les problèmes rencontrés par Grefenstette et Teufel, puisque dans

66. Le fait qu'ils ne disposent que de peu d'outils d'analyse syntaxique contraint d'ailleurs Grefenstette et Teufel à définir une façon simplifiée pour reconnaître un complément d'objet. Ils considèrent que le complément d'objet d'un verbe est le nom principal (i.e. la tête) du groupe nominal situé immédiatement avant la première préposition située après le prédicat.

leurs travaux, le verbe support recherché apparaît parmi les candidats retenus, mais il n'est pas toujours en première position.

Dras s'appuie pour cela sur le fait que les verbes candidats à être supports d'une nominalisation forment un ensemble restreint [67]. Ils apparaissent donc de façon anormalement élevée en position de verbe dans les phrases où se trouve un déverbal. On regrettera que l'auteur n'ait pas détaillé les mesures utilisées pour déterminer les associations entre les noms déverbaux et les candidats verbes supports. Les résultats présentés semblent pourtant pertinents.

L'aspect le plus intéressant de ce travail est de combiner une approche de l'information locale – le verbe support apparaît fréquemment comme prédicat du nom déverbal étudié – avec une approche globale – le verbe support se retrouve fréquemment dans ce rôle pour d'autres déverbaux.

4 Bilan et perspectives

Faire une synthèse de l'ensemble des travaux que nous avons présentés n'est pas une chose facile. En effet, la plupart de ceux-ci sont développés pour répondre à un besoin précis ou une application particulière. On peut toutefois remarquer que tous se situent dans le cadre de l'étude du langage utilisé dans un domaine spécifique, que ce soit pour l'identifier, l'organiser ou pour illustrer une démarche plus générale. La nature même des travaux d'extraction de connaissances sur corpus contraint en effet ceux-ci à se concentrer sur une partie du lexique de la langue. Or les mots avec lesquels sont construites les terminologies forment généralement un ensemble de lexèmes dont on suppose qu'il peut être circonscrit et organisé à l'aide de relation lexicales.

L'intégration de travaux tels que ceux que nous avons exposés dans une démarche générale de construction automatique d'un lexique reste cependant, à notre connaissance, à accomplir. Il est toutefois possible de citer quelques pistes, ce que nous nous sommes efforcés de faire dans cette étude. Un premier point important concerne l'adoption d'un formalisme satisfaisant pour décrire la structure du lexique et le contenu des entrées lexicales. Il faut également pouvoir circonscrire le vocabulaire autour duquel un tel lexique peut être construit, tâche réalisable grâce des outils d'extraction de terminologie.

67. Ce qui ne signifie pas que cet ensemble soit facile à circonscrire.

Un travail sur corpus doit alors permettre de mettre à jour les relations lexicales et sémantiques qui vont déterminer la structure du lexique, c'est-à-dire les relations entre les entrées lexicales. Enfin, il convient de trouver les informations qui permettront d'exprimer le sens attaché aux entrées lexicales. C'est ici que les travaux qui visent à déterminer quels sont les prédicats du domaine et comment ils expriment les relations au sein des entités syntaxiquement complexes du langage peuvent être utiles. Ce ne sont là, rappelons-le, que quelques idées.

Nous pouvons toutefois énoncer trois points qui doivent être étudiés avant tout travail d'acquisition lexicale sur corpus :

1. *les besoins :* il est important, pour des raisons évidentes, de préciser les besoins auxquels les connaissances lexicales obtenues doivent répondre ;

2. *les moyens :* il est tout aussi indispensable de savoir quels sont les moyens nécessaires à la mise en œuvre d'une méthode d'acquisition ; doit-on disposer d'une ébauche de lexique qu'on enrichira, d'un corpus étiqueté, etc. ? Il convient également de considérer attentivement ce qui est disponible afin d'éviter deux problèmes : redécouvrir des informations déjà connues, en particulier des éléments établis lors de la préparation manuelle des données nécessaires au fonctionnement de la méthode automatique d'acquisition, ou, au contraire, faire appel à des informations qui n'existent pas ;

3. *l'utilité :* enfin, toute tâche d'acquisition lexicale devant assister, voire, au moins partiellement, remplacer, le travail du lexicographe, il convient de réfléchir aux deux points précédents en terme de gain. Ainsi, une méthode qui obtient des résultats relativement simples mais dont les exigences sont modestes peut être considérée aussi utile qu'une autre méthode a priori plus performante mais qui nécessite un travail préalable très important. Les critères qui déterminent l'utilité d'une méthode sont à déterminer au cas par cas.

Références

[AB95] Houssem Assadi and Didier Bourigault. Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances. In *3èmes journées internationales d'analyse statistique de données textuelles*, Rome, Italie, décembre 1995.

[AB96] Houssem Assadi and Didier Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. In $10^{ème}$ *Congrès Reconnaissance des Formes et Intelligence Artificielle*, Rennes, France, janvier 1996.

[Aga95] Rajeev Agarwal. *Semantic Feature Extraction from Technical Texts with Limited Human Intervention.* PhD thesis, Mississippi State University, mai 1995.

[Ass96] Houssem Assadi. Interactive Semantic Analysis for Building Conceptual Models from Corpora. In *ECAI96*, Budapest, Hongrie, août 1996.

[Ass97] Houssem Assadi. Une méthode et des outils pour la construction d'une ontologie du domaine à partir de textes. In *JICAA 97*, Roscoff, France, mai 1997.

[Bac96] Bruno Bachimont. *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser.* PhD thesis, École polytechnique, mai 1996.

[BBC+87] Branimir Boguraev, Ted Briscoe, John Carroll, D. Carter, and C. Grover. The Derivation of a Grammatically-Indexed Lexicon from the Longman Dictionary of Comtemporary English. In 25^{th} *Meeting of the Association for Computational Linguistics*, Stanford, USA, 1987.

[BC94] Ted Briscoe and John Carroll. Toward Automatic Extraction of Argument Structure from Corpora. MLTT 06, Rank Xerox Research Center, Grenoble, France, 1994.

[BC97] Ted Briscoe and John Carroll. Automatic Extraction of Subcategorisation from Corpora. In 5^{th} *ACL conference on Applied Natural Language Processing*, Washington, USA, avril 1997.

[BHNZ97] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, and Pierre Zweigenbaum. Regroupements issus de dépendances syntaxiques en corpus: catégorisation et confrontation à deux modélisations conceptuelles. In *JICAA 97*, Roscoff, France, mai 1997.

[Bou94] Didier Bourigault. *Acquisition de terminologie.* PhD thesis, EHESS, 1994.

[BP96] Branimir Boguraev and James Pustejovsky, editors. *Corpus Processing for Lexical Acquisition.* MIT, 1996.

[BPV92] Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. Combining NLP and Statistical Techniques for Lexical Acquisition. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language Learning*, 1992.

[Bri94] Ted Briscoe. Parsing (with) Punctuation etc. MLTT 02, Rank Xerox Research Center, Grenoble, France, août 1994.

[Bui97] Paul Buitelaar. A Lexicon for Underspecified Semantic Tagging. In *ANLP97 Workshop on Tagging text with Lexical Semantics*, Washington, USA, avril 1997.

[Cru86] D. A. Cruse. *Lexical Semantics.* Cambridge Textbooks in Linguistics, 1986.

[Dai94] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, Université Paris VII, février 1994.

[Dow79] David Dowty. *Word Meaning and Montague Grammar.* D. Reidel Publishing Company, 1979.

[Dra95] Mark Dras. Automatic Identification of Support Verbs: A Step Towards a Definition of Semantic Weight. In *8th Australian Joint Conference on Artificial Intelligence*, 1995.

[Fab96] Cécile Fabre. *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations.* PhD thesis, Université de Rennes 1, décembre 1996.

[FS95] Cécile Fabre and Pascale Sébillot. Calculability of the Semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-Based Semantic Information. Rapport de recherche n°2742, INRIA, décembre 1995.

[GG95] Nicola Guarino and Pierdaniele Giaretta. *Towards Very Large Knowledge Bases*, chapter Ontologies and Knowledge Bases. Toward a Terminological Clarification. IOS Press, Amsterdam, 1995.

[GJ95] D. Garcia and A. Jackiewicz. Aide à l'acquisition des connaissances causales par exploration de textes. In *JAVA-95*, Grenoble, France, avril 1995.

[GMM94] R. Grishman, C. Macleod, and A. Meyers. Comlex Syntax: Building a Computational Lexicon. In *COLING*, Kyoto, Japon, 1994.

[Gre93a] Gregory Grefenstette. Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques. MLTT 01, Rank Xerox Research Center, Grenoble, France, septembre 1993.

[Gre93b] Gregory Grefenstette. Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. In *Workshop on Acquisition of Lexical Knowledge from Text. SIGLEX/ACL*, Columbus, USA, juin 1993.

[Gre94] Gregory Grefenstette. Corpus-Derived First, Second and Third-Order Word Affinities. In *EURALEX'94*, 1994.

[GT95] Gregory Grefenstette and Simone Teufel. Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations. In

7^{th} *Conference of European Chapter of the Association for Computational Linguistics*, Dublin, Irlande, mars 1995.

[Jac90] Ray Jackendoff. *Semantic Structures*. MIT, 1990.

[Jac97] Christian Jacquemin. Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Habilitation à diriger des recherches, Université de Nantes, janvier 1997.

[Lev93] Beth Levin. *English Verb Classes and Alternations*. University of Chicago Press, 1993.

[Man93] Christopher D. Manning. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In 31^{st} *Meeting of the Association for Computational Linguistics*, 1993.

[MBF+90] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University, juillet 1990.

[Mor97] Emmanuel Morin. Extraction de liens sémantiques entre termes dans des corpus de textes techniques : application à l'hyponymie. In *TALN97*, Grenoble, France, juin 1997.

[MPRT97] Chantal Muller, Xavier Polanco, Jean Royauté, and Yannick Toussaint. Acquisition et structuration des connaissances en corpus : éléments méthodologiques. Rapport de recherche n^o3198, INRIA, juin 1997.

[PAB93] James Pustejovsky, Peter Anick, and Sabine Bergler. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics*, 19(2), 1993.

[Pic96] Ronan Pichon. Typage des associations préférentielles dans les composés extraits de corpus. Mémoire de DEA, Université de Rennes 1, France, septembre 1996.

[Pot92] Bernard Pottier. *Sémantique Générale*. PUF, 1992.

[PS96] Victor Poznański and Antonio Sanfilippo. *Corpus Processing for Lexical Acquisition*, chapter Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora. MIT, 1996.

[PSDM94] Florence Pugeault, Patrick Saint-Dizier, and Marie-Gaëlle Monteil. Knowledge Extraction from Texts: a Method for Extracting Predicate-Argument Structures from Texts. In *COLING*, Kyoto, Japon, 1994.

[Pus91] James Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4), 1991.

[Pus95] James Pustejovsky. *The Generative Lexicon*. MIT Press, 1995.

[Rad88] Andrew Radford. *Transformational Grammar*. Cambridge Textbooks in Linguistics, 1988.

[Ras95] François Rastier. Le terme : entre ontologie et linguistique. *La banque des mots*, 7:35–65, 1995.

[Ras96] François Rastier. *Sémantique Interprétative*. PUF, 1996.

[RCA94] François Rastier, Marc Cavazza, and Anne Abeillé. *Sémantique pour l'analyse*. Masson, 1994.

[Res93] Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, décembre 1993.

[Res95a] Philip Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In 3^{rd} *Workshop on Very Large Corpora*, Cambridge, USA, juin 1995.

[Res95b] Philip Resnik. Using Information to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI*, Montréal, Canada, août 1995.

[Ril93] Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. In 10^{th} *National Conference on Artificial Intelligence (AAAI 93)*, 1993.

[Ril96a] Ellen Riloff. Automaticaly Generating Extraction Patterns from Untagged Text. In 13^{th} *National Conference on Artificial Intelligence (AAAI 96)*, Portland, Canada, août 1996.

[Ril96b] Ellen Riloff. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *AI journal*, 85, août 1996.

[RS97] Ellen Riloff and Jessica Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. In *Empirical Methods in Natural Language Processing*, Providence, USA, août 1997.

[Séb96] Pascale Sébillot. A Corpus-Based Approach to Refining a Domain-Independent Compound Interpretation System. In *NeMLaP-2*, Ankara, Turquie, 1996.

[VFP91] Paola Velardi, Michela Fasolo, and Maria Teresa Pazienza. How to Encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition. *Computational Linguistics*, 17(2), 1991.

[Weh97] Éric Wehrli. *L'Analyse Syntaxique des Langues Naturelles*. Masson, 1997.

[ZBHN97] Pierre Zweigenbaum, Jacques Bouaud, Benoît Habert, and Adeline Nazarenko. Coopération apprentissage en corpus et connaissance du domaine pour la construction d'ontologies. In *Journées Scientifiques et Techniques*, Avignon, France, avril 1997. Francil.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399